Statistical Methods for Comparative Effectiveness Research of Medical Devices

Statistical Methods for Comparative Effectiveness Research of Medical Devices

A dissertation presented

by

Lauren Margaret Kunz

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

October 2014

Dissertation Advisor: Professor Sharon-Lise T. Normand Lauren Margaret Kunz

Statistical Methods for Comparative Effectiveness Research of Medical Devices

Abstract

A recent focus in health care policy is on comparative effectiveness of treatments–from drugs to behavioral interventions to medical devices. Medical devices bring a unique set of challenges for comparative effectiveness research. In this dissertation, I develop statistical methods for comparative effectiveness estimation and illustrate the methodology in the context of three different medical devices. In chapter 2, I review approaches for causal inference in the context of observational cohort studies, utilizing a potential outcomes framework demonstrated using data for patients undergoing revascularization surgery with radial versus femoral artery access. Propensity score methods; G-computation; augmented inverse probability of treatment weighting; and targeted maximum likelihood estimation are implemented and their causal and statistical assumptions evaluated. In chapter 3, I undertake a theoretical and simulation-based assessment of differential follow-up information per treatment arm on inference in meta-analysis where applied researchers commonly assume similar follow-up duration across treatment groups. When applied to the implantation of cardiovascular resynchronization therapies to examine comparative survival, only 3 of 8 studies report arm-specific follow-up. I derive the bias of the rate ratio for an individual study using the number of deaths and total patients per arm and show that the bias can be large, even for modest violations of the assumption that follow-up is the same in the two arms. Furthermore, when pooling multiple studies with Bayesian methods for random effects meta-analysis, the direction and magnitude of the bias is unpredictable. In chapter 4, I examine the statistical power for designing a study of devices when it is difficult to blind patients and providers, everyone wants the

device, and clustering by hospitals where the devices are implanted needs to be taken into account. In these situations, a stepped wedge design (SWD) cluster randomized design may be used to rigorously assess the roll-out of novel devices. I determine the exact asymptotic theoretical power using Romberg integration over cluster random effects to calculate power in a two-treatment, binary outcome SWD. Over a range of design parameters, the exact method is from 9% to 2.4 times more efficient than designs based on the existing method.

Contents

List of Figures

List of Tables

Acknowledgments

Although only my name appears on the front of this dissertation, so many others have contributed to its production. My advisor, Dr. Sharon-Lise T. Normand, knew when to provide theorems, tissues, and tough love and helped guide me through this process. Next, I would like to thank my committee members, Dr. Francesca Dominici and Dr. Miguel Hernán. My collaboration with Dr. Donna Spiegelman has been a "step"(ped wedge) in a new and exciting direction and I look forward to continuing our work together.

I am grateful for the mentorship of Dr. Nancy Geller. She is a professional and personal inspiration. My brilliant classmates, turned friends, kept a sense of humor when I lost mine over coding errors and never ending problem sets. I would like to give special thanks to Mark Meyer and Allison Meisner Burke, who have provided an infinite amount statistical and moral support over the past years. Sheila Lee, Lindsey French, Jennie Gappa, Jinette Gappa Lais, Kevin Tatro, Sam McCabe, Ben Flink–distance cannot separate friends. Your visits provided moments to enjoy this city for more than my daily M2 rides across the Charles passing back and forth between Cambridge and Boston.

Finally, my entire family–especially Tom, Linda, and Kristin–provided unwavering support. Reminders of their pride give me more sense of accomplishment than letters after my name.

1. Introduction

A recent focus in health care policy is on comparative effectiveness of treatments–from drugs to behavioral interventions to medical devices. The demand for rigorous demonstrations of comparative effectiveness has led to previously developed statistical tools being utilized in new settings. Medical devices bring a unique set of challenges for comparative effectiveness research. After the introduction of medical devices in the 1950s and 1960s, increasing medical device technology, such as the development of cardiac pacemakers and prosthetic heart valves, prompted the FDA to propose a different approval process than that established for drugs. Medical device development and clinical assessment differs from drugs in many ways (Konstam et al. (2003)). A device evolves progressively, through refinement of its components and/or systems. For example, modification of an existing medical device may only involve a change in material or in a component, whereas improving a drug may involve combining two agents or a different biological target. The clinical effect of an implantable device is dependent upon the skill of the implanting physician, the so-called *learning curve effect*, whereas the clinical effect of a drug is not dependent on the skill of the prescriber. Study design issues, such as blinding and the aforementioned learning curve effects are challenging when assessing the effectiveness of a device compared to best medical therapy. Devices are often designed to perform multiple tasks and are not specifically engineered for one biologic target. Finally, devices are frequently designed to be used in conjunction with medications. These represent just some of the considerations when thinking about statistical methods for assessing the safety and effectiveness of medical devices in a comparative effectiveness setting.

This dissertation develops statistical methodology for comparative effectiveness assessments, including design considerations, of medical devices. In Chapter 2, I review the assumptions underpinning a causal analysis, linking to the potential outcomes framework developed by Rubin (Rubin (1974)). Methodology for binary treatments and a single outcome in the absence of randomization are reviewed. I discuss the causal and statistical assumptions associated with estimators based on propensity score matching, stratification and weighting; G-computation; augmented inverse probability of treatment weighting;

and targeted maximum likelihood estimation. A comparative assessment of the effectiveness of two different artery access strategies for patients undergoing percutaneous coronary interventions with a coronary stent illustrate the different approaches. Rudimentary R code is provided to assist the reader in implementing the various approaches. Like many inferential problems, some assumptions are not testable – for causal inference, these include the explicit assumption of potential outcomes, stable unit treatment value assignment (SUTVA), and ignorability of treatment assignment. In the artery access example, we find that all methods indicated a lower risk of in-hospital complications for the radial artery approach compared to the femoral approach, with the risk of in-hospital complications being approximately 1.6% lower in the radial group.

In Chapter 3, I undertake a theoretical and simulation-based assessment of the effect of differential follow-up information per treatment arm on inference in meta-analysis where the most common approach in clinical applications assumes follow-up duration is similar across treatment groups. The research is motivated by an investigation of the effectiveness of cardiac resynchronization therapy devices compared to those with cardioverter-defibrillator capacity where 3 of 8 studies report arm-specific follow-up duration. I derive the bias of the rate ratio when incorrectly assuming equal follow-up duration in the single study binary treatment setting. Simulations illustrate bias, efficiency, and coverage, and demonstrate that bias can be large, even for modest violations of the assumption that follow-up is the same in the two arms of an individual study. Combining study rate ratios with hierarchical Poisson regression models, I examine bias and coverage for the overall rate ratio via simulation in three cases: when average arm-specific follow-up duration is available for all studies, some studies, and no study. In the null case, bias and coverage are poor when the study average follow-up is used and improve even if some arm-specific follow-up information is available. As the rate ratio gets further from the null, bias and coverage remain poor. Furthermore, when pooling multiple studies with Bayesian methods for random effects meta-analysis, the direction and magnitude of the bias is unpredictable. When all studies are randomized trials, the impact of differential

follow-up is less likely to be an issue, as trials are designed to have equal follow-up in each arm.

In Chapter 4, I determine power for a binary treatment on a binary outcome in a cross-over cluster randomized design, referred to as a *stepped wedge cluster randomized design*. The design is motivated by the potential for large center effects in clinical trials of implantable medical devices and where the demand for the new device is high. Approximate power for a binary outcome based on a linear mixed model assuming normal variance has been proposed (Hussey and Hughes (2007)). Using maximum likelihood theory, I determine the exact asymptotic theoretical power for a two-tailed Wald test by capitalizing on computational advances using Romberg integration over the distribution of the cluster random effects. Power is compared among several designs, as well as to that found by Hussey and Hughes. I find that our method has higher power for the same design taking the binary nature of the outcome into account versus Hussey and Hughes. I use this method to design a study powered to detect effectiveness of a new left ventricular assist device (LVAD) model for patients with end-stage heart disease.

2. An Overview of Statistical Approaches for Comparative Effectiveness Research for Assessing In-Hospital Complications of Percutaneous Coronary Interventions By Access Site

Lauren M. Kunz[1], Sherri Rose[2], Donna Spiegelman[1,3], and Sharon-Lise T. Normand[1,2]

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Care Policy, Harvard Medical School
[3]Department of Epidemiology, Harvard School of Public Health

2.1 Introduction

Comparative effectiveness research (CER) is designed to inform health-care decisions by providing evidence on the effectiveness, benefits, and harms of different treatment options (AHR (2014)). While the typology of CER studies is broad, this chapter focuses on CER conducted using prospective or retrospective observational cohort studies where participants are not randomized to an intervention, treatment, or policy. We assume outcomes and covariates are measured for all subjects and there is no missing outcome or covariate information throughout; we also assume that the data are sampled from the target population – the population of all individuals for which the treatment may be considered for its intended purpose. Without loss of generality, we use the terms control or comparator interchangeably and focus on one non-time varying treatment. The scope of methods considered are limited to linear models – a single treatment assignment mechanism model and a single linear outcome model.

An example involving the in-hospital complications of radial artery access compared to femoral artery access in patients undergoing percutaneous coronary interventions (PCI) illustrate ideas. Coronary artery disease can be treated by a PCI in which either a balloon catheter or a coronary stent is used to push the plaque against the walls of the blocked artery. Access to the coronary arteries via the smaller radial artery in the wrist, rather than the femoral artery in the groin, requires a smaller hole and may therefore, reduce access-site bleeding, patient discomfort, and other vascular complications. Table 2.1 summarizes information for nearly 40,000 adults undergoing PCI in all non-federal hospitals located in Massachusetts. The data are prospectively collected by trained hospital data managers utilizing a standardized collection tool, sent electronically to a data coordinating center, and adjudicated (Mauri et al. (2008)). Baseline covariates measured include age, sex, race, health insurance information, comorbidities, cardiac presentation, and medications given prior to the PCI. Overall, radial artery access (new strategy) compared to femoral artery access (standard strategy) is associated with fewer in-hospital vascular and bleed-

ing complications (0.69% vs 2.73%). However, there is significant treatment selection – healthier patients are more likely to undergo radial artery access compared to those undergoing femoral artery access. Patients associated with radial artery access have less diabetes, more prior congestive heart failure, more left main coronary artery disease, and more shock compared to those undergoing femoral artery access. The CER question is: *When performing PCI, does radial artery access cause fewer in-hospital complications compared to femoral artery access for patients with similar risk?*

The remainder of the chapter provides the main building blocks for answering CER questions in settings exemplified by the radial artery access example – a single outcome with two treatment options. We sometimes refer to the two treatment groups as treated and comparator, exposed and unexposed, or treated and control. Notation is next introduced and the statistical causal framework is described. We adopt a *potential outcomes* framework to causal inference (Holland (1986)). The underlying assumptions required for CER are discussed. We restrict our focus to several major classes of estimators, and note that we do not exhaustively include all possible estimators for our parameter of interest. Approaches for assessing the validity of the assumptions follow and methods are illustrated using the PCI data.

2.2 Causal Model Basics

Assume a population of N units indexed by i each with an outcome, Y_i. In the radial artery example, units are subjects, and $Y_i = 1$ if subject i had a complication after PCI and 0 otherwise. Assume a binary-valued treatment such that $T_i = 1$ if the patient received the new treatment (e.g., radial artery access) and 0 (e.g., femoral artery access) otherwise. Approaches for treatments assuming more than two values, *multi-valued* treatments, generalize from those based on binary-valued treatments (see Imbens (2000), Lu et al. (2001)). Variables that are not impacted by treatment level and occur prior to treatment assign-

Table 2.1: Population characteristics stratified by type of intervention. All entries are percentages with the exceptions of number of observations, age, and number of vessels with $> 70\%$ stenosis.

	Intervention	
	Radial	**Femoral**
No. of Observations	5192	35022
Demographics		
Mean Age [SD]	63 [12]	65 [12]
Female	25.3	29.8
Race		
White	89.6	89.4
Black	3.3	3.2
Hispanic	4.3	3.5
Other	2.8	3.9
Health Insurance		
Government	46	50.3
Commercial	4.8	13.4
Other	49.2	36.3
Comorbidities		
Diabetes	33.1	32.7
Prior Congestive Heart Failure	9.4	12.7
Prior PCI	32	34.3
Prior Myocardial Infarction (MI)	28.7	30.1
Prior Coronary Artery Bypass Surgery	8.4	15.7
Hypertension	79.6	80.7
Peripheral Vascular Disease	12.1	12.8
Smoker	24.8	23.1
Lung Disease	13.7	14.4
Cardiac Presentation		
Multi-vessel Disease	10.3	10.9
Number of Vessels $> 70\%$ stenosis	1.49	1.58
Left Main Disease	3.7	7.2
ST-segment elevated MI	38.9	42.6
Shock	0.44	1.8
Drugs Prior to Procedure		
Unfractionated Heparin	87.3	61.7
Low Molecular Weight Heparin	3.83	4.27
Thrombin	25.5	54.9
G2B3A Inhibitors	26.7	26.8
Platelet Aggregate Inhibitors	85.8	86.6
Aspirin	98.2	97.5
In-Hospital Complication, %	0.69	2.73

ment are referred to as covariates. Let X_i denote a vector of observed covariates, all measured prior to receipt of treatment. Notation is summarized in Table 2.2. Within X, some covariates may be *confounders*. Confounding occurs due to differences in the outcome between exposed and control populations even if there were no exposure. The covariates that create this imbalance are called confounders (Greenland and Robins (1986)). Another type of covariate is an *instrumental variable* that is independent of the outcome and correlated with the treatment (see Imbens and Angrist (1994)). Instrumental variables, when available, are used when important key confounders are unavailable; their use is not discussed here. In the radial artery example, X includes age, race, sex, health insurance information, and cardiac and non-cardiac comorbidities. Because there are two treatment levels, there are two potential outcomes for each subject (Sekhon (2008)). Only one of the two potential outcomes will be observed for a unit.

Table 2.2: Notation for the potential outcomes framework to causal inference

Notation	Definition
T_i	Binary treatment for unit i (1=treatment; 0=comparator)
Y_i	Observed outcome for unit i
Y_{0i}	Potential outcome for unit i if $T_i = 0$
Y_{1i}	Potential outcome for unit i if $T_i = 1$
X_i	Vector of pre-treatment measured covariates for person i
μ_T	$E_X(E(Y \mid T = t, X))$, marginal expected outcome under t
Δ	$\mu_1 - \mu_0$, causal parameter

2.2.1 Causal Parameters

The idea underpinning a causal effect involves comparing what the outcome for unit i would have been under the two treatments – the *potential outcomes*. Let Y_{1i} represent the outcome for unit i under $T_i = 1$ and Y_{0i} for $T_i = 0$. The causal effect of the treatment on the outcome for unit i can be defined in many ways. For instance, interest may center on an *absolute effect*, $\Delta_i = Y_{1i} - Y_{0i}$, the *relative effect* $\Delta_i = Y_{1i}/Y_{0i}$, or on some other function of the potential outcomes. The fundamental problem of causal inference is that we only observe the outcome under the actual treatment observed for unit i, $Y_i = Y_{0i}(1 - T_i) + Y_{1i}(T_i)$.

A variety of causal parameters are available with the choice dictated by the particular problem. We focus on the causal parameter on the difference scale, $\Delta = \mu_1 - \mu_0$, where μ_1 and μ_0 represent the true proportions of complications if all patients had undergone radial artery access and femoral artery access, respectively. The marginal mean outcome under treatment $T = t$ is defined as

$$\mu_T = E_X \left(E(Y \mid T = t, X) \right), \tag{2.1}$$

averaging over the distribution of X. The marginal expected outcome is found by examining the conditional outcome given a particular values of X and averaging the outcome over the distribution of all values of X. The parameter μ_T is useful when interest rests on assessing population interventions. If the treatment effect is constant or homogeneous, then the marginal parameter is no different from the conditional parameter.

The average treatment effect (ATE) is defined as

$$E[Y_1 - Y_0] = E_X \left(E[Y \mid T = 1, X = x] - E[Y \mid T = 0, X = x] \right) \tag{2.2}$$

and represents the expected difference in the effect of treatment on the outcome if subjects were randomly assigned to the two treatments. The ATE includes the effect on subjects for whom the treatment was not intended, and therefore may not be relevant in some policy evaluations (Heckman et al. (1997)). For example, to assess the impact of a food voucher program, interest rests on quantifying the effectiveness of the program for those individuals *who are likely to participate* in the program. In this case, the causal parameter of interest is the average effect of treatment on the treated (ATT)

$$E_X \left(E[Y \mid T = 1, X = x] - E[Y \mid T = 0, X = x] \mid T = 1 \right). \tag{2.3}$$

The ATT provides information regarding the expected change in the outcome for a randomly selected unit from the treatment group.

Which causal estimand is of interest depends on the context. When randomized, on average, the treated sample will not be systematically different from the control sample, and

the ATT will be equal to the ATE. Throughout this chapter we focus on the ATE as the causal estimand of interest because (1) both radial and femoral artery access are a valid strategy for all subjects undergoing PCI and (2) we wish to determine whether fewer complications would arise if everyone had undergone radial artery access rather than femoral artery access.

2.2.2 Underlying Causal Assumptions

If the following untestable assumptions are violated, the causal parameters defined can be estimated statistically but cannot be interpreted causally. We begin with the explicit assumption of potential outcomes. The ability to state the potential outcomes implies that although an individual receives a particular treatment, the individual could have received the other treatment, and hence has the potential outcomes under both treatment and comparison conditions.

Stable unit treatment value assignment (SUTVA): No interference and no variation in treatment

The stable unit treatment value assignment (SUTVA) consists of two parts: (1) no interference and (2) no variation in treatment. SUTVA is untestable and requires subject matter knowledge. The no interference assumption implies that the potential outcomes for a subject do not depend on treatment assignments of other subjects. In the radial artery example, we require that radial artery access in one subject does not impact the probability of an in-hospital complication in another subject. If a subject's potential outcomes depends on treatments received by others, then $Y_i(T_1, T_2, ..., T_N)$, indicating outcome for subject i depends on the treatment received by $T_1, T_2, \cdots, T_N$. SUTVA implies

$$Y_i(T_1, T_2, ..., T_N) = Y_i(T_i) = Y_{it}. \tag{2.4}$$

Under what circumstances would the assumption of no interference be violated? Con-

sider determining whether a new vaccine designed to prevent infectious diseases – because those who are vaccinated impact whether a person becomes infected, there will be interference. The radial artery access example may violate the no interference assumption when considering the *practice makes perfect* hypothesis. As physicians increase their skill in delivering a new technology, the less likely complications arise in subsequent uses, and the more likely the physician is to use the new technology. Conditioning physician random effects would make the no interference assumption reasonable.

The second part of SUTVA states that there are not multiple versions of the treatment (and of the comparator), or that the treatment is well defined and the same for each subject receiving it. In the radial artery access example, if different techniques are used to access the radial artery (or the femoral artery) by different clinicians, then the SUTVA is violated.

Ignorability of treatment assignment

The most common criticism of CER using observational cohort studies involves the unmeasured confounder problem – the assertion that an unmeasured variable is confounding the relationship between treatment and the outcome. Ignorability of the treatment assignment or *unconfoundedness* of the treatment assignment with the outcome assumes that conditional on observed covariates, the probability of treatment assignment does not depend on the potential outcomes. Hence, treatment is effectively randomized conditional on observed baseline covariates. This assumption is untestable and can be strong, requiring observation of all variables that affect both outcomes and treatment in order to ensure

$$(Y_0, Y_1) \perp T \mid X \text{ and } P(T = 1 \mid Y_0, Y_1, X) = P(T = 1 \mid X). \tag{2.5}$$

2.2.3 Key Statistical Assumptions

Positivity

Positivity requires units at every combination of observed covariates,

$$0 < P(T = 1 \mid \mathbf{X}) < 1. \tag{2.6}$$

Structural violations of positivity occur when units associated with a certain set of covariates cannot possibly receive the treatment or control. The ATE and the ATT cannot be identified under structural violations of positivity. A treatment for use only in women, for example, requires exclusion of males. Practical violations of the positivity assumption may arise due to finite sample sizes. With a large number of covariates, there may not be subjects receiving treatment and control in strata induced by the covariate space. Positivity is a statistically testable assumption.

Constant treatment effect

A constant treatment effect conditional on X implies that for any two subjects having the same values of covariates, their observable treatment effects should be similar

$$\Delta_i \mid X = \Delta_j \mid X \;\; i \neq j. \tag{2.7}$$

Under a constant treatment effect, the ATE may be interpreted both marginally and conditionally. While this assumption can be empirically assessed, guidelines regarding exploratory and confirmatory approaches to determination of non-constant treatment effects should be consulted (see pco (2013)). Moreover, methods that have the average causal effect in the population as the estimand do not need to make any assumptions about constant treatment effect within levels of the confounders. These methods include IPTW, G-computation, and TMLE (see below).

2.3 Approaches

Under the assumptions described above, various approaches exist to estimate the ATE. The approaches are divided into three types: methods that model only the treatment assignment mechanism via regression, methods that model only the outcome via regression, and methods that use both the treatment assignment mechanism and outcome. Formulae are provided and general guidelines for implementation based on existing theory to assist the reader in deciding how best to estimate the ATE are described.

2.3.1 Methods Using the Treatment Assignment Mechanism

Rosenbaum and Rubin (Rosenbaum and Rubin (1983)) defined the propensity score as the probability of treatment conditional on observed baseline covariates, $e(X_i) = P(T_i = 1 \mid X_i)$. The propensity score, $e(X)$, is a type of balancing score such that the treatment and covariates are conditionally independent given the score, $T \perp X \mid e(X)$ so that for a given propensity score, treatment assignment is random. The true propensity score in observational studies is unknown and must be estimated. Because of the large number of covariates required to satisfy the treatment ignorability assumption, the propensity score is typically estimated parametrically by regressing the covariates on treatment status and obtaining the estimated propensity score, $\widehat{e(X)}$. Machine learning methods have been developed for prediction and have been applied to estimation of the propensity score (see Lee et al. (2009), McCaffrey et al. (2004), Setoguchi et al. (2008), van der Laan and Rose (2011)). Variables included in the propensity score model consist of confounders and those related to the outcome but not to the treatment. The latter are included to decrease the variance of the estimated treatment effect (Rubin (2007)). Instrumental variables, those related to treatment but not to the outcome should be excluded (Brookhart et al. (2006)). The rationale for the exclusion of instrumental variables under treatment ignorability relates to the fact that their inclusion does not decrease the bias of the esti-

mated treatment effect but does increase the variance. By their construction, propensity scores reduce the dimensionality of the covariate space so that they can be utilized to match, stratify, or weight observations. These techniques are next described. Inclusion of the propensity score as a predictor in a regression model of the outcome to replace the individual covariates constitutes a simpler dimension reduction approach compared to other estimators that use both the complete outcome regression and treatment mechanism (see Section 2.3.3). However, if the distribution of propensity scores differ between treatment groups, there will not be balance (Stuart (2010)) between treated and control units when using $\widehat{e(X)}$ as a covariate, subsequent results may display substantial bias (Kang and Schafer (2007)). Thus methods that do not make use of the propensity score, such as G-computation (Section 2.3.2) still benefit from an analysis of the propensity score, including testing for empirical violations of the positivity assumption.

Matching

Matching methods seek to find units with different levels of the treatment but having similar levels of the covariates. Matching based on the propensity score facilitates the matching problem through dimension reduction. Several choices must be made that impact the degree of incomplete matching (inability to find a control unit to match to a treated unit) and inexact matching (incomparability between treated and control units). These considerations include determination of the structure of the matches (one treated matched to one control, one-to-k, or one-to-variable), the method of finding matches (greedy versus optimal matching), and the closeness of the matches (will any match do or should only close matches be acceptable). The literature on matching is broad on these topics. We refer the reader to Rassen 2012 et al. (Rassen et al. (2012)) for discussion about matching structure, Gu and Rosenbaum (Gu and Rosenbaum (1993)) for a discussion on the comparative performance of algorithms to find matches, and Rosenbaum (Rosenbaum (2002)) for options for defining closeness of matches.

Let $j_m(i)$ represent the index of the unit that is mth closest to unit i among units with the opposite treatment to that of unit i, $\mathcal{J}_M(i)$ the set of indices for the first M matches for unit i, such that $\mathcal{J}_M(i) = j_1(i), ..., j_M(i)$, and $K_M(i)$ the number of times unit i is used as a match. Lastly, define $K_M(i) = \sum_{l=1}^{N} Ii \in \mathcal{J}_M(l)$ where I is the indicator function. Then the ATE estimator and its corresponding variance are

$$\begin{aligned} \hat{\Delta}_{Matching} &= \frac{1}{N}\sum_{i=1}^{N}\left[(2T_i - 1)\left(1 + \frac{K_M(i)}{M}\right)Y_i\right], \\ \text{Var}(\hat{\Delta}_{Matching}) &= \frac{1}{N^2}\sum_{i=1}^{N}\left(1 + \frac{K_M(i)}{M}\right)^2 \hat{\sigma}^2(X_i, T_i) \end{aligned} \tag{2.8}$$

where the conditional variance $\hat{\sigma}^2(X_i, T_i)$ is estimated as $\frac{J}{J+1}\left(Y_i - \frac{1}{J}\sum_{m=1}^{J} Y_{l_j(i)}\right)^2$ and J is a fixed number of observations. This approach (Abadie and Imbens (2006)) is implemented in the Matching package in the R software system. The variance formula does not account for estimation of the propensity score, only the uncertainty of the matching procedure itself. While adjustment for the matched sets in computing standard errors is debated (Stuart (2010)), we recommend that this design features be accounted for in the analysis.

Much of the preceding discussion assumed a larger pool of controls to find matches for treated subjects – estimators based using this strategy provides inference for the ATT. Estimating the ATE additionally requires identification of treatment matches for each control group unit. Therefore, the entire matching process is repeated to identify matches for units in the control group. The matches found by both procedures are combined and used to compute the ATE.

Stratification

Stratification methods, also referred to as sub-classification methods, divide subjects into strata based on the estimated propensity score. Within each stratum, treatment assignment is assumed random. As with matching, sub-classification can be accomplished

without using the propensity score, but this runs into problems of dimensionality. Commonly subjects are divided into groups by quintiles of the estimated propensity score, as Rosenbaum and Rubin (Rosenbaum and Rubin (1984)) showed that using quintiles of the propensity score to stratify eliminates approximately 90% of the bias due to measured confounders in estimating the absolute treatment effect parameter, $\Delta = Y_1 - Y_0$. The average effect is estimated in each stratum as the average of the differences in outcomes between the treated and control:

$$\hat{\Delta}_q = \frac{1}{N_{1q}} \sum_{i \in T \cap I_q} Y_i - \frac{1}{N_{0q}} \sum_{i \in C \cap I_q} Y_i$$

where N_{iq} is the number of units in stratum q with treatment i, I_q is indicates membership in stratum q, so $T \cap I_q$ would indicate that a subject in stratum q received the treatment. The overall average is computed by averaging the within strata estimates based on their sample sizes:

$$\begin{aligned} \hat{\Delta}_{Stratification} &= \sum_{q=1}^{Q} W_q \Delta_q; \quad W_q = \frac{N_{1q} + N_{0q}}{N} \\ \text{Var}(\hat{\Delta}_{Stratification}) &= \Sigma_q W_q^2 v_q^2; \quad v_q^2 = \frac{v_{1q}^2 + v_{0q}^2}{2} \end{aligned} \tag{2.9}$$

where $v_{iq}^2 = s_{iq}^2 / N_{iq}$. Because individuals in each stratum do not have identical propensity scores, there may be residual confounding (see Austin and Mamdani (Austin and Mamdani (2006))) and balance between treated and control units requires examination within strata.

Inverse Probability of Treatment Weighted Estimators (IPTW)

The intuition behind weighting is that units that are underrepresented in one of the treatment groups are up weighted and units that are overrepresented are down weighted. The ATE can be estimated as

$$\hat{\Delta}_{HT-IPTW} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i Y_i}{\widehat{e(X_i)}} - \frac{1}{N} \sum_{i=1}^{N} \frac{(1 - T_i) Y_i}{1 - \widehat{e(X_i)}} \tag{2.10}$$

using the estimated propensity score, $\widehat{e(X)}$. We denote this estimate HT-IPTW to acknowledge the Horvitz-Thompson (Horvitz and Thompson (1952)) ratio estimator utilized in survey sampling. IPTW estimators solve an estimating equation that sets the estimating function to zero and aims to find an estimator that is a solution of the equation. For example, consider

$$\sum_{i=1}^{N} D(\hat{\Delta})(T_i, Y_i, X_i) = 0,$$

with $D(\Delta)(T_i, Y_i, X_i)$ defining the estimating function and $\hat{\Delta}$ is an estimator of the parameter that is a solution of the estimating equation. Robins et al (Robins et al. (1995)) derived variance estimators, but bootstrapping can also be used. Inverse propensity score weighting is sensitive to outliers. Treated subjects with a propensity score close to one or control subjects with a propensity score close to zero will result in large weights. The weights can be trimmed but doing so introduces bias in estimation of the treatment effect (Potter (1993)). Robins, Hernan and Brumback (Robins et al. (2000)) propose using stabilizing weights, such that

$$\hat{\Delta}_{S-IPTW} = \left(\sum_{i=1}^{N} \frac{T_i}{\widehat{e(X_i)}}\right)^{-1} \sum_{i=1}^{N} \frac{T_i Y_i}{\widehat{e(X_i)}} - \left(\sum_{i=1}^{N} \frac{1-T_i}{1-\widehat{e(X_i)}}\right)^{-1} \sum_{i=1}^{N} \frac{(1-T_i)Y_i}{1-\widehat{e(X_i)}} \tag{2.11}$$

IPTW estimators are known to have problems with large variance estimates in finite samples. Inverse probability weights can be used to estimate parameters defined by a marginal structural model, which we do not discuss here (see Robins et al. (2000)). See Lunceford and Davidian (Lunceford and Davidian (2004)) and Stefanski and Boos (Stefanski and Boos (2002)) for details regarding deriving variances using the empirical sandwich method. Alternatively, a bootstrap procedure may be applied to the whole process including estimation of the propensity score.

2.3.2 Methods Using the Outcome Regression

Multivariable Regression Modeling

The ATE can be estimated by the treatment coefficient from regression of the outcome on the treatment and all of the confounders. The functional form of the relationship between the outcome and covariates needs to be correctly specified. The risk difference can be validly estimated by fitting a ordinary least squares regression model and using the robust variance to account for non-normality of the error terms. This approach is exactly equivalent to fitting a generalized linear model for a binomial outcome with the identity link and robust variance.

In the case of no overlap of the observed covariates between treatment groups, the model cannot be fit as the design matrix will be singular. Therefore, erroneous causal inferences are prohibited by the mechanics of the estimation procedure in the case of complete non-overlap. However, standardized differences should still be looked at to see how the treated and control groups differ, even under the assumption of no unmeasured confounding. If there is little overlap, we do not want to extrapolate to areas where we may not be justified in making causal inference.

G-Computation

G-computation (G-computation algorithm formula, G-formula, Generalized-computation) is completely non-parametric (Robins (1986)), but we focus on parametric G-computation, which is a maximum-likelihood-based substitution estimator (Snowden et al. (2011)). Substitution estimators involve using a maximum-likelihood-type estimator (e.g., regression, super learning, etc.) for the outcome regression and plugging it into the parameter mapping that defines the feature we are interested in estimating – here, that feature is the average treatment effect $\mu_1 - \mu_0$. Under ignorability of the treatment

assignment, the G-computation formula permits identification of the distribution of potential outcomes based on the observed data distribution. In step 1, a regression model or other consistent estimator for the relationship of the outcome with treatment (and covariates) is obtained. In step 2, (a) set each unit's treatment indicator to T=1 and obtain predicted outcomes using the fit from step 1 and (b) repeat step 2(a) by setting each unit's treatment indicator to T=0. The treatment effect is the difference between $\hat{Y}_{1i}$ and $\hat{Y}_{0i}$ for each unit, averaged across all subjects. When there are no treatment covariate interactions, linear regression and G-computation that uses a parametric linear regression provide the same answer for a continuous outcome. We can define this as

$$\hat{\Delta}_{G-comp} = \frac{1}{N}\sum_{i=1}^{N}[\hat{E}(Y \mid T_i = 1, X_i) - \hat{E}(Y \mid T_i = 0, X_i)] \tag{2.12}$$

where $\hat{E}(Y \mid T_i = t, X_i)$ is the regression of Y on X in treatment group $T = t$. Two points are worth noting. First, if the outcome regression is not estimated consistently, the G-computation estimator may be biased. Second, while positivity violations will not be obvious when implementing a G-computation estimator, they remain important to assess, and can lead to a non-identifiable parameter or substantially biased and inefficient estimate.

2.3.3 Methods Using the Treatment Assignment Mechanism and the Outcome

Double robust methods use both an estimator for the outcome regression and for the treatment assignment. Estimators in this class may be preferable because they are consistent for the causal parameters if either the outcome regression or treatment assignment regression are consistently estimated (Robins et al. (1994)). Two double robust methods include the augmented inverse probability of treatment weighted estimator (A-IPTW) and the targeted maximum likelihood estimator (TMLE).

Augmented Inverse Probability Weighted Estimators (A-IPTW)

Like IPTW estimators, A-IPTW estimators are also based on estimating equations but differ in that A-IPTW estimators are based on the *efficient influence curve*. An efficient influence curve is the derivative of the log-likelihood function with respect to the parameter of interest. The efficient influence curve is a function of the model and the parameter, and provides double robust estimators with many of their desirable properties, including consistency and efficiency (van der Laan and Robins (2003)). The A-IPTW for the ATE is

$$\begin{aligned}\hat{\Delta}_{A-IPTW} &= \frac{1}{N}\sum_{i=1}^{N}\frac{[I(T_i=1)-I(T_i=0)]}{\widehat{e(X_i)}}(Y_i-\hat{E}(Y\mid T_i,X_i)) \\ &+ \frac{1}{N}\sum_{i=1}^{N}(\hat{E}(Y\mid T_i=1,X_i)-\hat{E}(Y\mid T_i=0,X_i)) \end{aligned} \tag{2.13}$$

where $\hat{E}(Y \mid T_i = t, X_i)$ is the regression of Y on X in treatment group $T = t$, and $I()$ is an indicator function. The nuisance parameters in the estimating equation for the A-IPTW are the treatment assignment mechanism and the outcome regression. Further discussion of estimating equations and efficient influence curve theory can be found in (van der Laan and Rose (2011), van der Laan and Rubin (2006), van der Laan and Robins (2003)). Of note, A-IPTW estimators ignore the constraints imposed by the model by not being substitution estimators. For example, an A-IPTW estimator for a binary outcome many produce predicted probabilities outside the range [0,1]. Thus finite sample efficiency may be impacted, even though asymptotic efficiency occurs if both the outcome regression and treatment assignment mechanism are consistently estimated.

Targeted Maximum Likelihood Estimator (TMLE)

The TMLE has a distinct algorithm for estimation of the parameter of interest, sharing the double robustness properties of the A-IPTW estimator, but boasting additional statistical properties. TMLE is a substitution estimator, thus unlike the A-IPTW, it does respect the

global constraints of the model. Therefore, among other advantages, this improves the finite sample performance of the TMLE.

The TMLE algorithm for the ATE involves two steps. First, the outcome regression $E[Y \mid T, X]$ and the treatment assignment mechanism $e(X)$ are estimated. Denote the initial estimate $\hat{E}[Y \mid T, X] = \widehat{Q^0}(T, X)$ and the updated estimate

$$\widehat{Q^1}(T, X) = \widehat{Q^0}(T, X) + \hat{\epsilon} \left(\frac{T}{\widehat{e(X)}} - \frac{1 - T}{1 - \widehat{e(X)}} \right),$$

where ϵ is estimated from the regression of Y on $\frac{T}{\widehat{e(X)}} - \frac{1-T}{1-\widehat{e(X)}}$ with an offset $\widehat{Q^0}(T, X)$. The estimator for the ATE is the given by

$$\hat{\Delta}_{TMLE} = \frac{1}{N} \sum_{i=1}^{N} (\widehat{Q^1}(T = 1, X_i) - \widehat{Q^1}(T = 0, X_i)) \tag{2.14}$$

A parametric regression can be used to estimate both the outcome regression and the treatment assignment mechanism. However, the targeted learning framework allows for the use of machine learning methods to estimate these components in an effort to achieve consistent estimators (van der Laan et al. (2007), van der Laan and Rose (2011), Rose (2013)). Confidence intervals for both the A-IPTW and TMLE can be constructed using influence curve methods or bootstrapping techniques(van der Laan and Rose (2011),van der Laan and Rubin (2006),van der Laan and Robins (2003)).

2.4 Assessing Validity of Assumptions

2.4.1 Ignorability

Ignorability of the treatment assignment is not directly testable and largely assessed by subject matter knowledge. Several strategies can bolster the viability of the assumption(Rosenbaum (1987)) however. Multiple control or comparison groups that differ with respect to an unmeasured confounder, if available, can be used. If outcomes between the

two control groups do not differ, then this observation would support the argument that the unmeasured confounder is not responsible for any treatment-control outcome differences. Another option is to identify an outcome that is associated with an unmeasured covariate but where a treatment would not expect to have any effect. Such outcomes, referred to as *control* outcomes, provide a means to detect unobserved confounding. Tchetgen (Tchetgen (2014)) proposes a method to correct estimates using control outcomes. Finally, Rosenbaum (Rosenbaum (2002)) provides approaches to perform a sensitivity analyses for an unobserved confounder through examination of a range of potential correlations between the unobserved confounder and the treatment assignment, and the unmeasured confounder and the outcome.

2.4.2 Positivity

Positivity or overlap can be measured through examination of the distributions of covariates for the treated and control subjects. While there are many measures of balance, the difference in average covariates scaled by the sample standard deviation, d, provides an intuitive metric. It is calculated as

$$d = \frac{\bar{x}_{1j} - \bar{x}_{0j}}{\sqrt{\frac{s_{1j}^2 + s_{0j}^2}{2}}} \tag{2.15}$$

where $\bar{x}_{ij}$ is the mean of covariate j among those with treatment i and s_{ij} is the estimated standard deviation. The quantity d is interpreted as the number of standard deviations the treated group is above the control group. Mapping the standardized differences to percentiles provides a mechanism to describe the extent of non-overlap between two groups. For instance, a standardized difference of 0.1 indicates 7.7% non-overlap of the two normal distributions; a standardized difference of 0 indicates complete overlap of the two groups; and a standardized difference of 0.7 corresponds to 43.0% non-overlap. Rules of thumb suggest that a standardized difference less than 0.1 is negligible Normand et al.

(2001)). Examination of the standardized differences alone characterizes only marginal distributions – the distribution of individual covariates. Because areas of weak overlap may exist, reviewing the distributions of the estimated propensity scores stratified by treatment groups is recommended.

2.4.3 Constant treatment effect

The assumption of a constant treatment effect may be explored by introducing interactions between the treatment and sub-group indicators; or by dividing the population into subgroups based on X_i, estimating an average causal effect within each subgroup, and comparing the constancy of subgroup specific causal effects. Cases in which the treatment effect may not be constant should be identified a priori as well as the size of meaningful treatment effect heterogeneity in order to avoid multiple testing.

2.5 Radial Versus Femoral Artery Access for PCI

We return to the PCI example introduced earlier to determine whether access via the radial artery reduces the risk of in-hospital complications compared to access via the femoral artery. Table 2.1 indicates imbalances between the radial and femoral artery accessed subjects. For instance, the standardized difference for use of thrombin is -62.85% indicating 40% non-overlap between the distribution of thrombin use for those undergoing PCI via the radial artery and those via the femoral artery. Ten of the observed covariates have percent standardized differences greater than 10.

2.5.1 Estimating Treatment Assignment: Probability of Radial-Artery Access

The propensity score was estimated using logistic regression. The set of covariates initially considered were determined by conversations with cardiologists who perform PCI. A primary model specification was selected for determining radial versus femoral artery access and included all covariates linearly as well as interactions (see the Appendix). Visual examination of a density plot of the estimated linear propensity scores by treatment arm (Figure 2.1) provides insights into the observable differences between the radial and femoral groups. The overlap assumption may also tested with a formal comparison test, such as the Kolmogorov-Smirnov nonparametric test. If there is little overlap, excluding subjects with extreme propensity score values may be necessary. For example, there are some subjects undergoing femoral artery accessed PCI whose estimated linear propensity scores do not overlap with the linear propensity scores for radial artery accessed subjects (Figure 2.1, density to the left of values of -5.0). Dropping subjects will make the estimates only valid for the region of common support. For the PCI example, visual inspection indicates that there are no patients who have a very high probability of radial treatment (on the probability scale, values near 1). The majority of subjects in both groups have low propensity scores, but those receiving radial artery access have higher propensity scores on average, as expected.

2.5.2 Approaches

Using the estimators described earlier, we determine the comparative effectiveness of radial artery access relative to femoral artery access. For comparability among estimates, all 95% interval estimates reported below are constructed using robust standard errors (1000 bootstrap replicates or theoretical results).

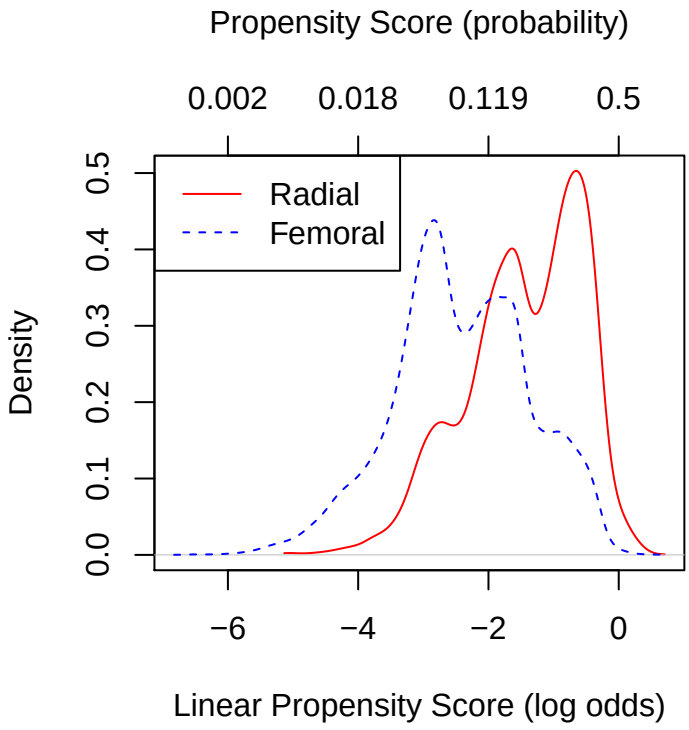


Figure 2.1: Density of estimated linear propensity scores, $\widehat{\text{logit}(e(X_i))}$, by artery access strategy. Larger values of the propensity score correspond to a higher likelihood of radial artery access. The upper horizontal axis gives the scale of the actual estimated probabilities of radial artery access.

Matching on the Propensity Score. Using the Matching program in R, we implement 1-1 matching without replacement to estimate the ATE. First, we identified femoral-artery accessed matches for each radial-artery accessed subject and next, found radial-artery accessed matches for each femoral-artery accessed subject. This resulted in 10326 matched pairs using a caliper of 0.2 standard deviations of the linear propensity score. The caliper was necessary in order to reduce the standardized differences for all covariates to below 0.1. In the matched sample, 42 of the radial artery subjects were used only once, 5142 were used twice, and 8 were not used; 7084 of the femoral artery subjects were used once, 1621 were used twice, and 26317 were not used. After matching, the percent standardized mean differences (Table 2.3 and Figure 2.2) improved. The linear propensity scores for radial artery and femoral artery accessed subjects in the matched sample overlap substantially (Figure 2.3).

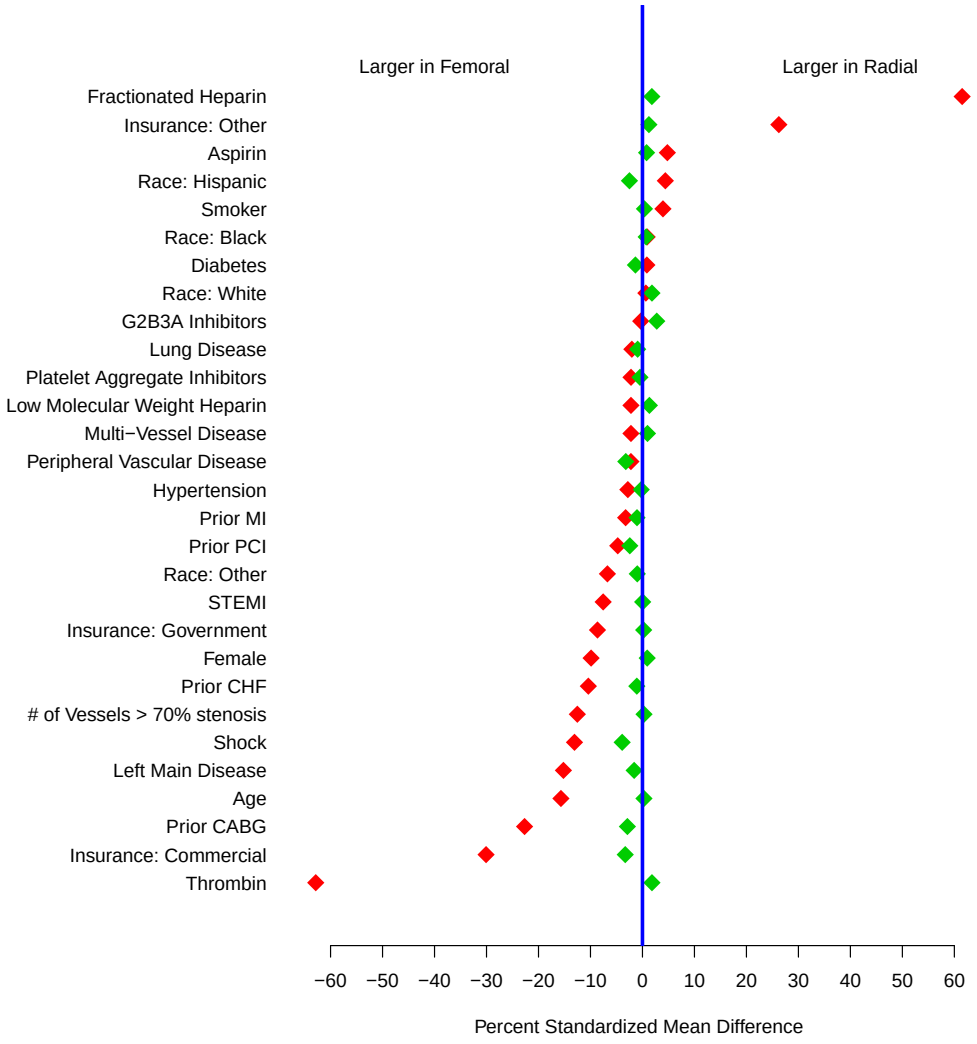


Figure 2.2: Percent standardized mean differences before (red) and after matching (green), ordered by largest positive percent standardized mean difference before matching.

The ATE estimated using matching and corresponding 95% confidence interval are

$$\hat{\Delta}_{Matching} = -0.0143(-0.0182, -0.0104) \tag{2.16}$$

indicating subjects undergoing PCI via radial artery access were 1.43% less likely to have an in-hospital complication compared to those accessed via the femoral artery. Using a more stringent caliper moved the point estimate further from the null, but discarded more

Table 2.3: Population characteristics pre and post matching listed by type of intervention. All are reported as percentages, except the number of procedures, age, and number of vessels. Positive standardized differences indicates a larger mean in the radial artery group.

	Pre-Match			Post Match		
	Intervention		% Standardized	Intervention		% Standardized
	Radial	Femoral	Mean Diff	Radial	Femoral	Mean Diff
No. of Procedures	5192	35022		10326	10326	
Demographics						
Mean Age [SD]	63 [12]	65 [12]	-15.68	63 [12]	63 [12]	0.29
Female	25.3	29.8	-9.88	25.4	25.0	0.91
Race						
White	89.6	89.4	0.66	89.7	89.2	1.83
Black	3.3	3.2	0.91	3.2	3.1	0.72
Hispanic	4.3	3.5	4.40	4.3	4.8	-2.50
Other	2.8	3.9	-6.74	2.7	2.9	-1.00
Health Insurance						
Government	46	50.3	-8.65	46.2	46.1	0.23
Commercial	4.8	13.4	-30.09	4.8	5.6	-3.31
Other	49.2	36.3	26.24	48.9	48.3	1.24
Comorbidities						
Diabetes	33.1	32.7	0.85	33.1	33.8	-1.35
Prior CHF	9.4	12.7	-10.41	9.5	9.8	-1.08
Prior PCI	32	34.3	-4.75	32.1	33.3	-2.44
Prior MI	28.7	30.1	-3.24	28.7	29.2	-1.05
Prior bypass surgery	8.4	15.7	-22.68	8.4	9.3	-2.90
Hypertension	79.6	80.7	-2.81	79.7	79.8	-0.24
Peripheral vascular disease	12.1	12.8	-2.22	12.1	13.2	-3.21
Smoker	24.8	23.1	3.94	24.9	24.7	0.38
Lung disease	13.7	14.4	-2.04	13.8	14.1	-0.92
Cardiac Presentation						
Multi-vessel Disease	10.3	10.9	-2.21	10.2	9.9	0.97
# of Vessels $>$ 70% stenosis	1.49	1.58	-12.53	1.49	1.49	0.30
Left main Disease	3.7	7.2	-15.21	3.8	4.1	-1.60
ST-elevated MI	38.9	42.6	-7.56	39.1	39.1	0.04
Shock	0.44	1.8	-13.08	0.4	0.7	-3.90
Drugs Prior to Procedure						
Heparin (unfractionated)	87.3	61.7	61.50	87.2	86.6	1.81
Heparin (low weight molecular)	3.83	4.27	-2.21	3.8	3.6	1.33
Thrombin	25.5	54.9	-62.85	25.7	24.9	1.83
G2B3A inhibitors	26.7	26.8	-0.33	26.8	25.6	2.75
Platelet Aggregate inhibitors	85.8	86.6	-2.20	86.2	86.4	-0.48
Aspirin	98.2	97.5	4.79	98.2	98.1	0.79
In-Hospital Complication, %	0.69	2.73		0.69	2.09	

observations.

We note two additional facts. First, the estimate of the ATT is -0.0145 (standard error = 0.0023), a slightly larger benefit in those likely to undergo PCI via the radial artery. Second, because we created matched pairs, McNemar's test could also be used for inference. The number of pairs in which the in-hospital complication rates differed within members of the pairs was 285 (2.76% of the 10326 matched pairs). Among the 285 discordant pairs, the number of pairs in which the radial artery accessed member had an in-hospital com-

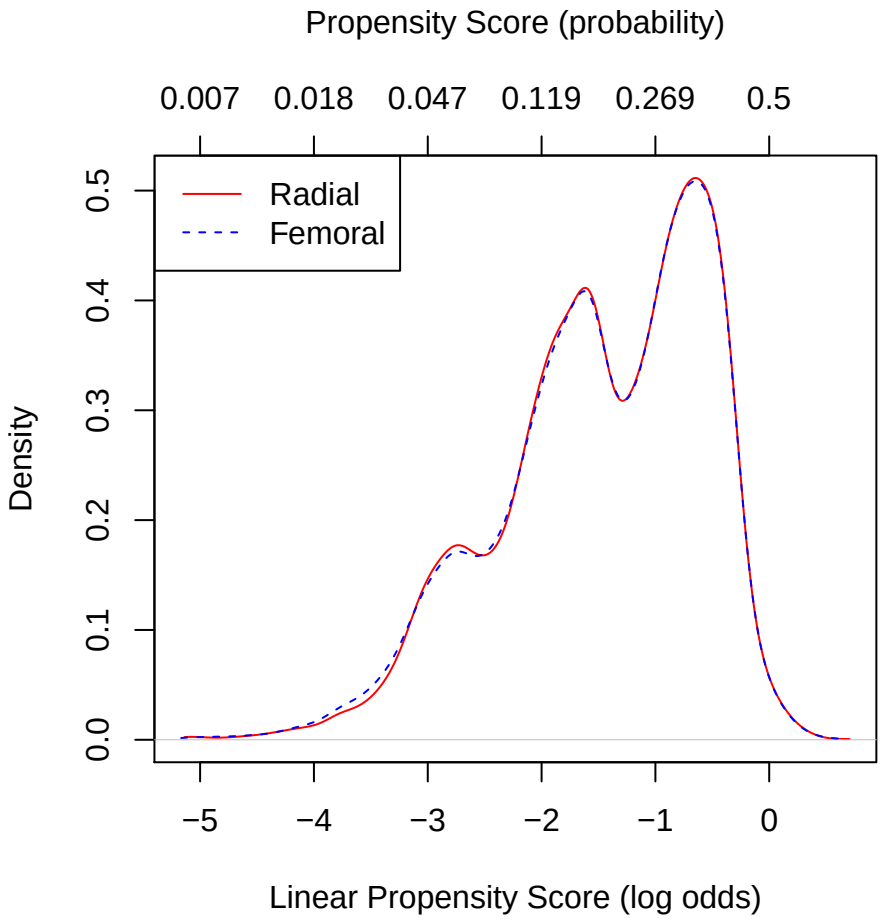


Figure 2.3: Density of estimated linear propensity scores, $\widehat{\text{logit}(e(X_i))}$, after matching by artery access strategy. Larger values of the propensity score correspond to a higher likelihood of radial artery access. The top axis gives the scale of the actual estimated probabilities of radial artery access.

plication was 70 (0.25 of discordant pairs). This value is lower than the null of 0.5 and indicates a benefit of radial artery access.

Stratification on the Propensity Score. The 40214 subjects were grouped into five strata using estimated propensity scores (Table 2.4). In the lowest quintile ($q = 1$), 2.46% of subjects fell into the radial artery access group whereas in the highest quintile, 31.8% of the subjects were accessed via the radial artery. If the propensity scores are balanced in each stratum, the covariates in each stratum should also be balanced. However, only

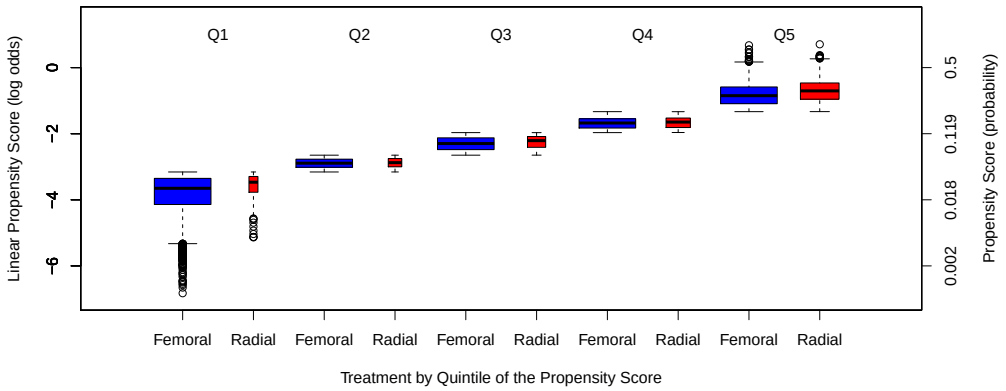


Figure 2.4: Boxplots of the linear propensity scores (log odds of radial artery access) by quintile. Boxplot widths are proportional to the square root of the samples sizes. The right axis gives the scale of the actual estimated probabilities of radial artery access.

using 5 strata did not result in balanced propensity scores for the PCI data. Two sample t-tests within strata showed significant differences in the propensity scores for the radial and femoral groups, although visually, the linear propensity scores appear quite similar within strata (Figure 2.4). There was less balance in the extreme quintiles, as is often the case.

Table 2.4: Properties of the quintiles based on the propensity score where $q = 1$ has the smallest values of the propensity score and $q = 5$ the largest. For each quintile, sample sizes and percentages of subjects undergoing radial artery access, the difference in mean in risk of complications ($\hat{\Delta}_q$, Section 2.3.1), and the average estimated propensity score are reported.

Stratum	**Radial**		**Femoral**		**Average**
q	N_{1q}	%	N_{0q}	$\bar{y}_{1q} - \bar{y}_{0q}$	$\widehat{e(X)}$
1	198	2.46	7845	-0.0135	0.0246
2	435	5.41	7608	-0.0147	0.0529
3	753	9.36	7289	-0.0217	0.0926
4	1249	15.53	6794	-0.0181	0.1588
5	2557	31.79	5486	-0.0158	0.3166
Overall	5192	12.91	35022	-0.0168	0.1291

The stratum-specific estimates are consistent – in every quintile, radial artery accessed patients were less likely to have complications compared to femoral artery accessed patients. Quintile-specific estimates were combined to obtain an overall $\hat{\Delta}_{Stratification}$= -

0.0168 (-0.0213, -0.0122) (Section 2.3.1) indicating that subjects undergoing PCI via radial artery access were 1.68% less likely to have in-hospital complications compared to those accessed via the femoral artery. Caution should be exercised in interpreting this estimate given the imbalance between treatment group still present within quintiles. Increasing the number of strata from 5 to 10 did not eliminate imbalance of the linear propensity scores between radial and femoral subjects within strata based on two sample t-tests (6 out of the 10 strata had p-values > 0.05). Achieving balance requires modifying the logistic model for treatment assignment or eliminating subjects residing in areas of non-overlap. In our example, we are unable to find a propensity score method that balances the data using stratification and thus conclude that this approach is unsuitable. Haviland et al found their treatment and control groups to be too dissimilar to warrant continued propensity score analysis (Haviland et al. (2007)).

Weighting by the Propensity Score. To implement weighting by the inverse probability, we estimate the weights as $1/\widehat{e(X)}$ for the radial subjects and $1/(1-\widehat{e(X)})$ for the femoral subjects. The weights are strongly right skewed having a maximum of 170 (radial artery accessed subject) and median of 1.11 leading to $\hat{\Delta}_{HT-IPTW} = -0.0168(-0.0214, -0.0122)$. The stabilized point and interval estimates are $\hat{\Delta}_{S-IPTW} = -0.0169(-0.0214, -0.0124)$. The results are similar, both indicating a benefit of radial artery access. However, the maximum weight, even after stabilizing, remained large with a value of 22.

Multivariate Regression. We estimate the ATE for a few different multivariate regression models with robust standard errors. Adjusting for all measured covariates using indicators for quintiles of age and number of vessels with $> 70\%$ stenosis, the ATE was $-0.0160(-0.0205, -0.0112)$. All potential confounders have events, so it is reasonable with our sample size to include all known and suspected risk factors in the multivariable model. A stepwise selected model with a liberal entry/exit criteria (p-value=0.2) was also run where when any variable where 1 or more levels were selected, all levels were forced into the final model. This model resulted in an ATE closer to the null value, $-0.0152(-0.0200, -0.0107)$.

G-Computation. G-computation does not use the model for the propensity score. To estimate the ATE using G-computation we assume a linear relationship between in-hospital complications and all covariates and the treatment indicator. The estimated coefficients and standard errors are reported in Table 2.5. The key parameter is the coefficient of the term *Radial* which is estimated as -0.016 (standard error = 0.002). Using all the estimated regression coefficients to obtain predictions and differencing yields $\hat{\Delta}_{G-comp} = -0.0160(-0.0189, -0.0127)$.

Table 2.5: Estimated coefficients (standard errors) of the outcome model.

	Estimated	
Covariate	**Coefficient**	**Standard Error**
Female	0.017	(0.002)
Diabetes	0.001	(0.002)
Smoker	0.001	(0.002)
Prior PCI	-0.004	(0.002)
Prior MI	0.0002	(0.002)
Prior CABG	-0.007	(0.003)
Prior CHF	0.027	(0.002)
Lung Disease	0.006	(0.002)
STEMI	0.012	(0.002)
Race: Black	-0.008	(0.004)
Race: Hispanic	-0.003	(0.004)
Race: Other	-0.0002	(0.005)
Insurance: Commercial	0.005	(0.003)
Insurance: Other	-0.002	(0.002)
Shock	0.068	(0.006)
Left Main Disease	0.022	(0.003)
Age	0.001	(0.0001)
Multi-Vessel Disease	0.008	(0.003)
# of Vessels $>$ 70% stenosis	-0.001	(0.001)
Peripheral Vascular Disease	0.006	(0.002)
Hypertension	0.003	(0.002)
Aspirin	0.0002	(0.005)
Fractionated Heparin	0.002	(0.002)
Low Molecular Weight Heparin	0.001	(0.004)
G2B3A Inhibitors	0.021	(0.002)
Platelet Aggregate Inhibitors	-0.006	(0.002)
Thrombin	-0.001	(0.002)
Radial Access	**-0.016**	**(0.002)**
Constant	-0.034	(0.008)

Augmented-IPTW. The augmented IPTW uses both the model for the outcome and the propensity score (see Section 2.5.1 and Appendix Table) for an estimate of $\hat{\Delta}_{AIPTW} = -0.0164(-0.0210, -0.0118)$. The risk of in-hospital complications is 1.64% lower in the

radial group.

Targeted Maximum Likelihood Estimation. To estimate the ATE using TMLE, we utilize the tmle package in R. We supply parametric models for the outcome and the propensity score (see Section 2.5.1 and Appendix Table) to compare results. The resulting treatment effect estimate is $\hat{\Delta}_{TMLE} = -0.0163(-0.0209, -0.0117)$. Inferences are similar to the earlier findings – a lower risk of in-hospital complications associated with radial compared to femoral artery access.

2.5.3 Comparison of Approaches

The results of the various approaches to estimation of the effectiveness of radial artery access compared to femoral artery access for PCI are similar (Figure 2.5). Each indicated a lower risk of in-hospital complications for the radial artery approach compared to the femoral approach. The only method that discarded subjects was matching on the propensity score which may explain why the estimated risk difference for this method differed from the others. The ATE based on G-computation had the shortest confidence interval (width = 0.62), 2/3 the size of the largest.

Table 2.6: Model Results: estimated coefficient of the treatment effect, radial versus femoral artery access on any in-hospital complications (robust standard errors).

	Estimated	
Method	**Coefficient**	**Standard Error**
Matching	-0.0143	(0.0020)
Stratification	-0.0168	(0.0023)
IPTW	-0.0168	(0.0023)
Multivariate regression	-0.0160	(0.0024)
G-computation	-0.0160	(0.0016)
A-IPTW	-0.0164	(0.0023)
TMLE	-0.0163	(0.0023)

Did we make reasonable assumptions? We indicated that the SUTVA may be violated as a consequence of (a) patients nested with physicians and (b) the practice makes perfect hy-

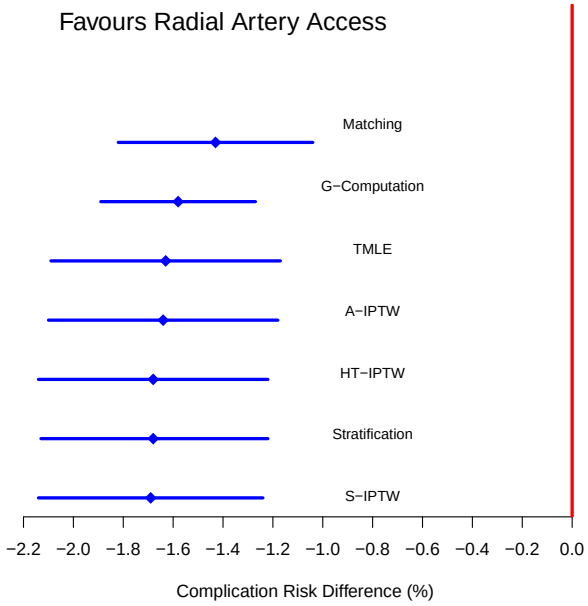


Figure 2.5: Comparison of results, ordered by size of ATE estimate. All methods use the same model for treatment assignment and outcome. All 95% confidence intervals are based on 1000 bootstrap replicates.

pothesis. A reasonable next step would involve the inclusion of random physician effects. The positivity assumption is met for the matching estimator because of the restrictions we placed on identifying matches. However, there may be regions of imbalance for the other estimators using the estimated propensity scores - we observed residual confounding when using the stratified estimate. In terms of treatment assignment ignorability, we do not have a control outcome nor an additional comparison group. We did determine that to attribute the ATE to an unobserved confounder rather than to radial artery access, an unobserved confounder would need to produce a 2.5 fold increase in the odds of radial access (beyond that adjusting for the set of covariates we have already included). Is it plausible that such a confounder exists? Of course one could exist – to place the size of the unmeasured confounder into context, the odds associated with fractionated heparin is 7.1 and with no shock of 3.9. Finally, there is an a-priori reason to believe that the effectiveness of radial versus femoral artery access may differ between males in females, that

is, there may be a non-constant treatment effect. Using matching estimators, we find a benefit of radial artery access in both males and females, with the ATE in females twice that of males: female ATE = -0.0267 (standard error = 0.004317) and male ATE = -0.0100 (standard error = 0.001583).

2.6 Concluding Remarks

As the need for comparative effectiveness research grows, reliance on observational cohort studies will as well. Several estimators are available to researchers to infer the effect of interventions. Causal inference can be made from observational studies whenever there is no bias. In cross-sectional studies, standard multivariable modeling methods can be used to obtain causal estimates, but the model needs to include all confounders properly parameterized. Double robust methods may be helpful if we have more knowledge about the model for treatment assignment than the model for the outcome. All involve statistical assumptions, and as we have described, some fundamental causal assumptions that are not testable. In this chapter we reviewed a selected set of estimators – our review is by no means comprehensive and are for cross-sectional data only. The reader is strongly encouraged to read the articles we have referenced. An example involving the choice of artery to utilize when unblocking clogged arteries illustrated the assumptions required, empirical evidence to support the assumptions when possible, and the estimates. The data involved nearly 40,000 subjects and the availability of many covariates. Even with this relatively small dataset, its dimensionality required reduction in order to facilitate analyses despite having a clear CER question. We focused on a single treatment assignment mechanism model and a single outcome model – clearly, more than one model may fit the data and meet the assumptions. As data acquisition technologies grow, researchers will be faced with making more analytical decisions when conducting empirical studies. These decisions should be made transparent to readers.

Acknowledgements

Dr. Normand's effort was supported by FDA/Chickasaw Nation Industries contract HHSF223201110172C (MDEpiNet Methodology Center) and by U01-FD004493 (MDEpiNet Medical Counter Measures Study). We gratefully acknowledge the Massachusetts Department of Public Health for permitting the use of the radial artery access example data.

3. Comparative Effectiveness and Meta-Analysis of Cardiac Resynchronization Therapy Devices: The Role of Differential Follow-up

Lauren M. Kunz[1], Sharon-Lise T. Normand[1,2], Danica Marinac-Dabic[3] and Art Sedrakyan[4]

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Care Policy, Harvard Medical School
[3]Division of Epidemiology, Center for Devices and Radiological Health, FDA
[4]Weill Cornell Medical College of Cornell University and New York Presbyterian Hospital

3.1 Introduction

A common statistical tool used to infer the comparative effectiveness of treatments is through a meta-analysis where study-specific estimates obtained from the literature are combined using standard statistical principles. With an increasing medical literature and wider range of statistical modeling techniques possible through computational advances, the types of meta-analyses have also increased (Sutton and Higgins (2008)). Researchers combine study effect estimates not only from randomized trials, but also from observational studies (Stroup et al. (2000)), from a combination of randomized and observational studies via cross-design synthesis (Droitcour et al. (1993)), and from trial arms corresponding to different studies via network meta-analysis (Lumley (2002)). The expanding reliance on network or cross-design meta-analyses highlights differential follow-up duration, which must be accounted for when events occur at any point over a follow-up period and censoring occurs throughout that period.

The conventional meta-analysis includes I primary studies using summary information, $\{Y_{ij}, n_{ij}\}$, such as a statistic and sample size, for each treatment arm j, about a parameter, θ_{ij}, with the common objective of making inferences about a population parameter, μ. When follow-up duration varies by treatment arm, we require the exposure in each study arm, $\bar{e}_j = \sum_{k=1}^{n_j} e_{jk}/n_j$, where e_{jk} is the follow-up for person k in arm j. While use of incident rate models or survival models is common within studies, most applied researchers continue to utilize odds ratios as the primary measures of associations, modeling probabilities of events within a particular time frame to combine study summaries. A BioMed Central article (Tierney et al. (2007)) reviewing one year of the Cochrane Library, for example, reported that the majority of cancer-related meta-analyses (63%) employed odds ratios or relative risks rather than hazard ratios.

When the follow-up time is fixed, modeling the probability of death is accomplished using the number of events and the total number of people, while assuming follow-up duration is the same or similar for the treatment arms. For meta-analyses of time to event

data, the log hazard ratio can be estimated directly if the observed number of events and the log rank expected number of events in each group for each study are reported, or if the log hazard ratio and its variance from the results of a Cox regression (Woods et al. (2010)) are available. Parmar (Parmar et al. (1998)) details other methods when the hazard ratio with confidence interval, p-value for the Mantel-Haenszel version of the log rank statistic, or when the published survival curves are given.

Differential follow-up duration in meta-analysis of observational studies is particularly challenging. In work for the Food and Drug Administration's Medical Device Epidemiology Network (MDEpiNet), we require assessing safety and effectiveness of cardiac resynchronization therapy (CRT) devices compared to CRT devices with cardioverter-defibrillator capacity (CRT-D). Both are implanted pacemakers used to improve mechanical synchrony in patients with heart failure and involve the placement of three leads (right and left atrium, and right ventricle). The CRT-D has an added defibrillation capability to break fast arrhythmias. The devices differ in costs (Feldman et al. (2005)) − average patient costs are higher for CRT-D ($82,200) compared to CRT-alone ($59,900). While there is a lack of clinical trials designed to assess the incremental benefit of CRT-D compared to CRT alone, the vast majority of patients receiving therapy with biventricular pacing are now implanted with CRT-D devices. Table 3.1 provides a summary of the 8 studies that compare CRT-D and CRT-alone. This comparison is part of a larger evidence synthesis project with interest in comparing CRT-D, CRT-alone, and optimal medical therapy. The search strategy was based on a previously published comprehensive review of cardiac resynchronization therapy and implantable cardioverter-defibrillators in left ventricular systolic dysfunction sponsored by the Agency for Healthcare Research and Quality.

A data synthesis of the CRT studies presents a number of complications. First, the average length of follow-up is 25.7 months with a standard deviation of 15.3 months across the studies. The constancy of the log hazard across this time frame is questionable and a full-accounting of the follow-up time is required. Second, not all primary studies use a survival time approach. The information available for all-cause mortality include the

number of deaths per arm, the total number of patients per arm, and the average length of follow-up across both arms (rather than average per arm). The typical approach for determining person-months of follow-up per arm involves multiplying the reported average months of follow-up time by the total number of people enrolled in each treatment group. The average all-cause mortality rate, across all primary studies, is 8.83 deaths per 1000 person-months: 7.37 deaths per 1000 person-months in the CRT-D arm and 10.63 deaths per 1000 person-months in the CRT-alone arm. For the observational studies the mortality rate is 8.43 per 1000 person-months, whereas for the RCT studies the mortality rate is higher, at 10.03 deaths per 1000 person-months. Table A.2 in the Appendix provides more follow-up information by study where the evidence suggests differential follow-up time by treatment arm. Only 3 studies provide any information regarding arm-specific follow-up time.

Using theoretical calculations in Section 3.2, we derive the bias of the rate ratio in the setting of a single study with two treatment groups. We use simulations to illustrate bias, efficiency, and coverage when ignoring variable treatment arm follow-up duration. We present a model to combine rate ratios in the meta-analytic setting. In this setting, we utilize simulation to characterize the operating characteristics of the estimators as a function of the duration of differential follow-up and describe how the availability of follow-up by treatment arm impacts these estimators. In Section 3.3 we perform a data analysis of the CRT-D and CRT studies. We close with recommendations for performing meta analyses when follow-up varies by treatment arm and when this information is unavailable in Section 3.4.

Table 3.1: CRT-D versus CRT-alone primary studies: All-cause mortality and other study summaries. IHD = ischemic heart disease; NYHA = New York Heart Association; LVEF = left ventricular ejection fraction; QRS represents the time it takes for depolarization of the ventricles. ? indicate that the data was not reported.

		No. of	Follow-up (months)		**All-Cause Mortality/N**	
Study	Year	Patients	CRT-D	CRT	CRT-D	CRT
Adlbrecht et al. (2009)	2009	205	?	?	19/110	9/95
Stabile et al. (2009)	2009	233	56.8	60.1	49/116	53/117
Bai et al. (2008)	2008	542	?	?	73/395	57/147
Auricchio et al. (2007)	2007	1298	?	?	91/726	119/572
Ermis et al. (2004)	2004	126	13	18	8/62	26/64
Pappone et al. (2003)	2003	135	?	?	6/88	9/47
TOTAL		2539	?	?	246/1497	273/1042
Bristow et al. (2004)	2004	1212	16	16.5	105/595	131/617
Schuchert et al. (2013)	2013	402	?	?	20/228	19/174
TOTAL		1614	?	?	125/823	150/791

	Baseline characteristics					
Study	Mean Age (years)	% Male	% IHD	% NYHA Class III	Mean % LVEF	Mean (SD) QRS (milliseconds)
Adlbrecht et al. (2009)	65	78	46	83	27.5	158 (31)
Stabile et al. (2009)	69	77	49	69	26.5	≤ 120
Bai et al. (2008)	67	77	67	81	20	162 (24)
Auricchio et al. (2007)	64	76	43	80	24	168 (29)
Ermis et al. (2004)	69	96	56	87	22	?
Pappone et al. (2003)	64	76	43	100	28	153 (11)
MEAN	66.3	80	50.7	83.3	24.7	160.3
Bristow et al. (2004)	67	67	55	87	21	160 (?)
Schuchert et al. (2013)	68	80	50	85	25	163 (?)
MEAN	67.5	73.5	52.5	86	23	161.5

3.2 Methods

3.2.1 A Single Study

Consider a two-arm study with interest centered on a rate ratio for a control ($j = 0$) and a treatment arm ($j = 1$). Assume the number of events from treatment arm j is $Y_j \sim \text{Pois}(\theta_j)$ where the expected number of events is θ_j. The average length of follow-up for the jth treatment arm is $\overline{e_j} = \sum_{k=1}^{n_j} e_{jk}/n_j$ where k indexes individuals and n_j indexes the total number of individuals in the jth treatment arm. We write $\theta_j = \lambda_j \times \overline{e_j} \times n_j$, with λ_j as the mortality rate defined as $\lambda_j = \xi \exp(\omega \times j)$. The parameter ξ represents the outcome rate in the $j = 0$ arm and ω is the log rate ratio of the outcome in the $j = 1$ arm compared to the $j = 0$ arm. The maximum likelihood estimator (MLE) of ω is

$$\hat{\omega} = \log\left(\frac{\hat{\lambda_1}}{\hat{\lambda_0}}\right) = \log\left(\frac{Y_1/\overline{e_1}n_1}{Y_0/\overline{e_0}n_0}\right) \tag{3.1}$$

because $\hat{\lambda_j} = \frac{Y_j}{\overline{e_j}n_j}$. When average follow-up is the same in each treatment arm the MLE depends only on the number of subjects in each arm, $\hat{\omega} = \log\left(\frac{Y_1/n_1}{Y_0/n_0}\right)$.

The most common approach utilized when follow-up information is not reported separately for each treatment arm but rather reported for the overall study, $\bar{e}$, is to assume follow-up duration is the same in each arm $\bar{e} = \bar{e}_1 = \bar{e}_0$ so that the estimator becomes:

$$\hat{\omega^*} = \log\left(\frac{\hat{\lambda_1^*}}{\hat{\lambda_0^*}}\right) = \log\left(\frac{Y_1/n_1}{Y_0/n_0}\right) \tag{3.2}$$

with $\hat{\lambda_j^*} = \frac{Y_j}{\bar{e}n_j}$. Both the correct and incorrect estimators for the rate ratio (RR), defined as

$\exp(\omega)$, are biased (see Appendix A.2.2 for derivations) with $f = \frac{\overline{e_1}}{\overline{e_0}} > 0$ such that

$$E\left\{\widehat{\exp(\omega)} - \exp(\omega)\right\} = \exp(\omega)\left[\frac{\xi\overline{e_0}n_0}{(\xi\overline{e_0}n_0)^2 + 3(\xi\overline{e_0}n_0) + 1}\right] \text{ and} \tag{3.3}$$

$$E\left\{\widehat{\exp(\omega^*)} - \exp(\omega)\right\} = \exp(\omega)\left[(f-1) + \frac{f\xi\overline{e_0}n_0}{(\xi\overline{e_0}n_0)^2 + 3(\xi\overline{e_0}n_0) + 1}\right]. \tag{3.4}$$

When $\bar{e}_1 = \bar{e}_0$, $f = 1$ so that the bias is identical. When $f < 1$, implying longer follow-up in the $j = 0$ arm, the term $(f - 1)$ in Equation 3.4 is negative so that the incorrect estimator underestimates the true rate ratio. When $f > 1$, a similar argument indicates $\widehat{RR^*}$ overestimates the true rate ratio.

Single Study Simulations

To illustrate the impact of unequal follow-up we conducted a simulation study using 1000 experiments under a variety of conditions. We assumed $\overline{e_0} = 24$ months for the CRT arm, varied follow-up in the $j = 1$ arm CRT-D arm using $\bar{e}_1 = f \times \bar{e}_0 = 24f$ and permitted f to range from 0.8 to 1.3 by 0.05 step increments. For instance, the values of f for the Stabile, Ermis, and COMPANION studies reported in Table A.2 are 0.945, 0.722, and 0.970 respectively. We assume there are an equal number of people in each arm $n_0 = n_1 = 200$. The baseline rate in the control CRT arm is taken to be $\xi = 0.01$ deaths per person-month. Results are presented in Figure 3.1 under 3 values of the rate ratio: RR=1, 0.7, and 0.5 (large difference).

The impact of using the study average follow-up rather than the arm-specific follow-up can be large. The simulations confirm the theoretical bias results. In general, when follow-up is shorter in the treated arm, $f < 1.0$, the incorrect method of estimating the rate ratio underestimates the true rate ratio, whereas if follow-up is longer in the treated arm, $f > 1.0$, we overestimate the true rate ratio. The MSE for the incorrect rate ratio is greater than the MSE for the correct rate ratio when $f \neq 1$. As f moves away from one, coverage for the incorrect rate ratio drops from the desired 95%. Moreover, while the bias is larger for all $f \neq 1.0$ with the incorrect estimator, the relative efficiency favors

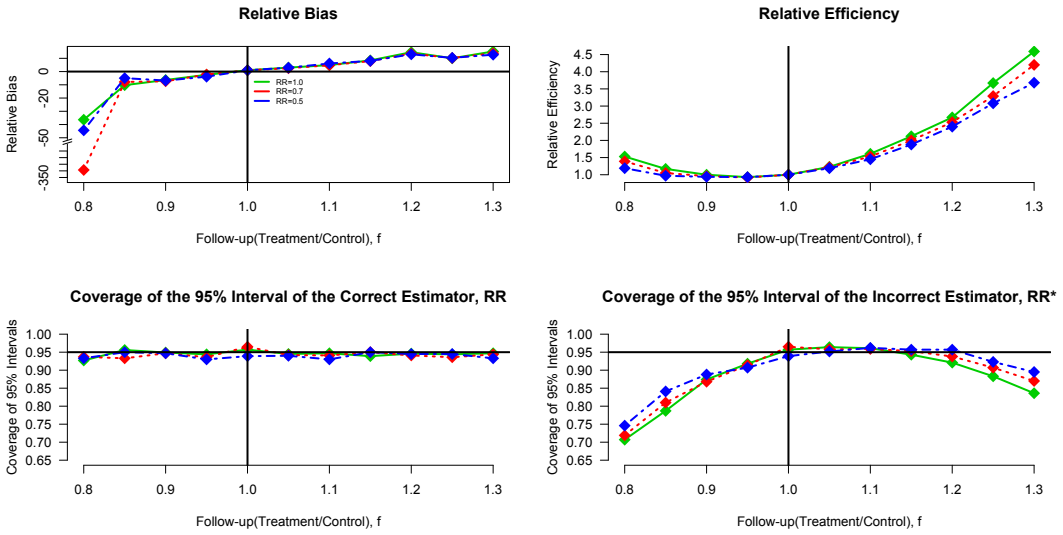


Figure 3.1: Simulation results for single study as function of relative follow-up in treatment arms: Each experimental condition is based on 1000 simulated datasets; $f = \frac{\overline{e_1}}{\overline{e_0}}$. Percent Bias = $\frac{\hat{\theta}-\theta}{\theta} \times 100$; RB = Relative Bias = Bias(RR*)/Bias(RR); MSE = Mean Squared Error=$1/1000 \times \sum(\hat{\theta} - \theta)^2$; and RE = Relative Efficiency = MSE(RR*)/MSE(RR).

the incorrect estimator when f is slightly less than 1 (around 0.9). We also examined experiments in which the sample sizes in the study arms were unequal and the same pattern of increasing relative bias away from $f = 1.0$ as in Figure 3.1 (results not shown).

3.2.2 Multiple Studies

Rather than one study, suppose there are I primary studies such that the number of events from arm j, study i, are $Y_{ij} \sim \text{Poisson}(\theta_{ij})$ with

$$\theta_{ij} = \lambda_{ij} \times \overline{e_{ij}} \times n_{ij} \text{ where } \lambda_{ij} = \xi_i \exp(\omega_i \times j). \tag{3.5}$$

As before, ξ_i is event rate for the $j = 0$ arm in the ith study; ω_i is the log rate ratio of the event for $j = 1$ versus $j = 0$ in the ith study; λ_{ij} is the event rate; n_{ij} is the total number of individuals; $\overline{e_{ij}}$ is the average person-months of follow-up; and θ_{ij} is the expected number of events. The baseline rate and the log relative rate ratio are assumed to vary

across studies to accommodate between-study variation using

$$\xi_i \overset{\text{indep.}}{\sim} \text{Gamma}(a, b) \text{ and } \omega_i \overset{\text{indep.}}{\sim} \text{Normal}(\mu, \sigma^2). \tag{3.6}$$

The selection of a Gamma distribution for the event rate in the control arm ensures positivity and is commonly used in hierarchical models with Poisson data (Carlin and Louis (2001)). The choice of a normal distribution for the log relative rate accommodates both positive and negative values.

A fully Bayesian approach places proper distributions for all the hyperparameters but we focus here on μ and σ^2. The model defined by equations (3.5) - (3.6) assumes that the $\overline{e_{ij}}$ are known for all i and j. When the average person-months of follow-up per arm and the total number of people per arm are available, then the total follow up by arm $\sum_{k=1}^{n_{ij}} e_{ijk} = n_{ij} \times \overline{e_{ij}}$. However, as before, we assume $\overline{e_{ij}} = \sum_{k=1}^{n_{ij}} e_{ijk}/n_{ij}$ are unavailable and rather $\overline{e_i} = \frac{\overline{e_{i1}} n_{i1} + \overline{e_{i0}} n_{i0}}{n_{i1} + n_{i0}}$ are reported. Primary interest remains focused on $\exp(\mu)$, the overall rate ratio across all studies. Because we are combining estimates across studies, σ the between-study standard deviation, is also a key parameter. In our motivating study, with such few primary studies, the overall results will be sensitive to this parameter.

Multiple Study Simulations

We generated 1000 experiments under 36 different parameter configurations: 3 relative rates $\times$ 2 standard deviation values $\times$ 6 values of f. We assume a moderate number of studies, I=20 studies. We fixed μ, the summary log rate ratio for the I studies, and σ, the between study standard deviation of the log rate ratios. Twenty individual log rate ratios, ω_i, were drawn from $\text{N}(\mu, \sigma)$. We also fixed the shape and scale parameters for simulating ξ_i from a Gamma distribution at $a = 2.55$ and $b = 0.00445$ implying a mean baseline rate of $a \times b = 1.13$ per 100 person-months and sampled 20 baseline rates in $j = 0$ arm. As in the single study simulations, we assume equal sample sizes in the two arms within each

study, but permitted the sample sizes across studies to vary. This was accomplished by drawing sample sizes from a Uniform(50,1000).

The average follow-up times in the control arms, $\overline{e_{i0}}$, were generated using a mixture of uniforms. We assumed 9 studies had $\overline{e_{i0}} \sim$ Uniform(10,32); 9 studies from Uniform(33, 57); and 2 studies from a Uniform(58,200). Under this scenario, between-study follow-up could range from 10 to 200 months. We let $f = \overline{e_{i1}}/\overline{e_{i0}}$ and generate average number of deaths $\theta_{ij} = \lambda_{ij} \times \overline{e_{ij}} \times n_{ij}$ and use this to simulate the number of deaths assuming the data are Poisson.

We estimated the person-time assuming full knowledge of the follow-up in both arms (e.g, the correct exposures) and then again using average exposure across both arms (e.g., the incorrect method). We fit the data using the WinBUGS software (Lunn et al. (2000)) and place non-informative priors for $\xi_i \sim$ Gamma$(2.5, 224.7)$ and $\omega_i \sim N(\mu, \sigma^2)$ where $\mu \sim N(0, 10^6)$ and $\sigma \sim$ Half-Normal(0.26). The choice of the prior distribution for the between-study standard deviation is much more influential in a hierarchical model than the choice of the prior distribution for the overall mean. The Half-Normal distribution ensures positivity of the standard deviation and because it has a mode at 0, also permits no differences between studies. A Half-Normal(0.26) indicates that 0.26 is the variance and yields a median value for σ of 0.39 with a 95% quantile of 1.0 for the between study standard deviation of the log rate ratio.

We ran one chain with 20000 iterations, 10000 burn in and thin every 10, resulting in 1000 Markov Chain Monte Carlo (MCMC) iterations. Convergence was assessed using the Geweke diagnostic in order to have a resulting 1000 simulations. Inference for the parameters is based on posterior means of the resulting 1000 MCMC iterations; these posterior means are averaged over the 1000 simulations. Credible intervals are found by taking the 2.5% and 97.5% percentiles of the 1000 iterations sorted. Coverage is found by calculating the frequency out of 1000 that the true value of the parameter falls between the 2.5% and 97.5% percentiles.

As f moves away from 1.0 the bias and coverage are worse for the incorrect RR (Table 3.2). Bias and coverage of σ are not impacted by f, and are similar for the correct and incorrect methods (Table 3.2). The relative bias (the percent bias of the incorrect estimator divided by the percent bias of the correct estimator) tends to be larger in magnitude moving away from the null value $RR = 1$. The relative bias is larger in magnitude when σ^2 is smaller (results not shown).

Partially Reported Follow-up Times

To address the problem of partially reported follow-up duration information by study arm, we examined two situations. In the first, we assume that some studies do not report follow-up by study arm completely at random (MCAR). In the second, the "missingness" mechanism is at random (MAR) and is related to whether the study is observational or randomized. All experimental conditions used in the previous multiple study simulations hold. For the MCAR case, we assume there is 17.5% missingness on average implying the total number of studies missing follow-up by treatment arm varies by simulation, but is usually 3 or 4 studies. For the MAR case, we assume that there is 5% missingness for randomized studies and 30% missingness for observational studies. We randomly made 10 of the 20 studies observational. The models are fit using the same fully Bayesian Poisson model in WinBUGS as previously described. When the simulated experiment reported follow-up by study arm, that information was utilized and when arm-specific follow-up was unavailable, the average study follow-up was used.

In general, use of partially observed follow-up times has a bias for the rate ratio that is between the bias for the correct case of having complete arm-specific follow-up and the incorrect case of having incomplete arm-specific follow-up for all primary studies (Table A.3). Coverage of the rate ratio is similar for both cases of missingness. As f moves away from 1.0, the bias and coverage worsen for the RR under both MAR and MCAR, but are more pronounced when σ^2 is larger (Figure 3.2). Bias and coverage of σ do not follow any

Table 3.2: Bias and coverage of the rate ratio, $\exp(\mu)$, and between-study standard deviation, σ, using partially reported follow-up times: Simulation results for 20 primary studies as a function of relative follow-up in treatment arms. Percent bias [(estimated - true)/true$\times$ 100].

	RR=1				RR=0.7				RR=0.5			
	$\sigma^2 = 0.01$		$\sigma^2 = 0.05$		$\sigma^2 = 0.01$		$\sigma^2 = 0.05$		$\sigma^2 = 0.01$		$\sigma^2 = 0.05$	
f	Corr	Incorr	Corr	Incorr	Corr	Incorr	Corr	Incorr	Corr	Incorr	Corr	Incorr
RR: **Bias**												
0.9	2.00	-8.07	5.32	-4.67	-0.86	-10.49	3.33	-6.84	3.34	-6.58	5.38	-4.82
0.95	3.22	-1.68	0.62	-4.51	1.16	-4.17	3.40	-1.54	-4.20	-9.34	-1.86	-6.70
1.0	2.51	2.52	-0.87	-0.84	2.54	2.54	-4.67	-4.67	0.78	0.78	-0.62	-0.60
1.05	-0.66	4.57	1.87	7.15	2.71	8.41	2.69	8.11	-1.54	3.50	4.16	9.42
1.1	-0.13	9.87	-2.79	6.58	2.71	13.00	-2.91	6.46	-0.58	9.02	-0.14	9.34
1.2	-1.96	17.16	4.49	24.81	-1.01	18.79	-1.10	18.37	1.42	21.58	5.16	25.72
RR: **Coverage**												
0.9	0.986	0.322	0.987	0.995	0.998	0.044	0.998	0.970	0.956	0.774	0.986	**0.996**
0.95	0.951	0.990	1.000	0.978	0.990	0.940	0.994	0.991	0.955	0.447	0.992	0.819
1.0	0.963	0.963	1.000	1.000	0.983	0.984	0.998	0.990	0.995	0.994	0.998	0.999
1.05	0.991	0.844	1.000	0.750	0.993	0.651	0.999	0.902	0.986	0.968	0.998	0.897
1.1	0.997	0.346	0.989	0.931	0.973	0.201	1.000	0.992	1.000	0.636	1.000	0.879
1.2	0.984	0.000	0.991	0.001	0.991	0.002	1.000	0.115	0.982	0.000	0.989	0.001
σ: **Bias**												
0.9	9.90	15.60	12.75	12.75	10.30	10.05	18.92	17.89	47.60	46.30	5.50	6.62
0.95	9.80	10.90	-8.45	-9.66	39.40	42.20	-11.76	-10.73	61.30	63.50	-24.51	-25.00
1.0	-4.80	-4.60	-7.02	-6.98	24.90	25.10	10.64	10.60	48.10	48.10	-9.97	-9.93
1.05	-1.10	-2.90	-34.75	-34.39	53.30	57.40	-3.40	-3.53	50.90	50.90	11.81	12.16
1.1	20.10	17.30	-13.15	-14.22	52.10	52.50	28.17	27.82	52.80	55.20	11.23	9.62
1.2	16.50	14.10	-8.36	-7.11	32.20	30.90	7.83	8.32	24.70	23.80	-0.98	-0.89
σ: **Coverage**												
0.9	0.991	0.985	0.994	0.992	0.988	0.989	0.983	0.986	0.869	0.885	0.998	0.998
0.95	0.985	0.984	0.996	0.994	0.892	0.872	0.987	0.991	0.684	0.656	0.890	0.885
1.0	0.983	0.980	0.995	0.996	0.937	0.937	0.989	0.990	0.849	0.842	0.981	0.975
1.05	0.988	0.991	0.650	0.667	0.774	0.731	1.000	1.000	0.842	0.847	0.994	0.991
1.1	0.969	0.978	0.962	0.958	0.754	0.757	0.942	0.947	0.807	0.783	0.998	0.998
1.2	0.980	0.979	0.999	1.000	0.956	0.960	0.999	0.998	0.956	0.961	0.999	1.000

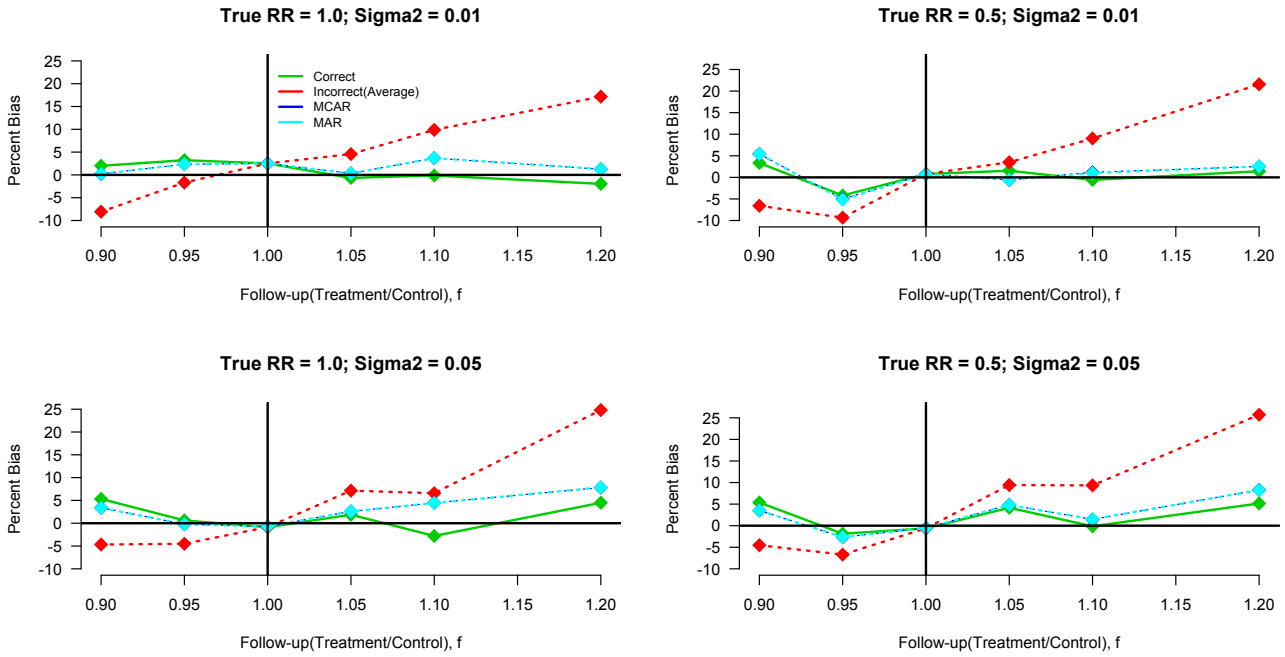


Figure 3.2: Percent Bias for the overall rate ratio via simulation in four cases for various RR and σ^2: arm-specific follow-up is available for all studies (correct), some studies (with "missingness" at random (MAR) and completely at random (MCAR)), and no study (average).

trend related to f and again are similar for both types of missingness (Table A.3).

3.3 Data Analysis: Effectiveness of CRT-D vs CRT

We analyzed the mortality data reported in Table 3.1 using the model described in Equations (3.5) - (3.6). Because of the small number of studies for the analysis, we considered a total of nine different sets of prior distributions for μ (the overall log rate ratio) and σ (the between-study standard deviation) that ranged in terms of informativeness. Models were estimated using the WinBUGS software and ran until convergence as determined by the Geweke score for the between-study standard deviation component. The all-cause mortality rate averaged across all primary studies is 8.83 deaths per 1000 person-months (for the CRT-D arm the mortality rate is 7.37 deaths per 1000 person-months and for the CRT alone arm the mortality rate is 10.63 deaths per 1000 person-months). Seven of the 8

studies show a benefit of CRT-D over CRT alone with a rate ratio less than 1.

Only 3 of the 8 studies report arm-specific follow up, with f =0.72 Ermis et al. (2004), 0.95 Stabile et al. (2009), and 0.97 Bristow et al. (2004). We estimate the overall rate ratio using two approaches: in the first, we use arm-specific follow-up when available and the study average when it is not (as in Section 3.2.2). In the second, we ignore any arm-specific follow-up information and use the study average follow-up.

3.3.1 Prior Distributions

We assumed the overall log rate ratio arose from a normal distribution centered at the null value of 0 (a rate ratio of 1.0). We selected three different variances for μ (overall log rate ratio): (1) the variance is 2 yielding a 95% interval from -2.77 to 2.77 (0.063 to 15.96 on the rate ratio scale) which is vague; (2) variance is 10 indicating the 95% interval for log rate ratio could range -6.2 to 6.2 which is quite vague; and (3) a variance of 1000000 which is extremely vague.

Three different prior distributions for the between-study standard deviation were selected. Two half-normal distributions permitted the underlying log rate ratio for a study to (1) have a median value of 0.39 with 95% quantile of 1.0 (Half-Normal(0.26)) and (2) have a median value of 0.14 with 95% quantile of 0.36 (Half-Normal(0.03)). A uniform distribution (Uniform(0,0.7)) had a mean and median of 0.35.

3.3.2 Results

When using the arm-specific follow-up information (for 3 of the 8 studies that provided it), the posterior mean (95% credible interval) of the overall rate ratio was 0.71 (0.49, 0.96) and 0.71 (0.55, 0.89) under the most non-informative pair of priors, $\mu \sim$ Normal(0,1e06) and $\sigma \sim$ Half-Normal(0.26), and most informative pair of priors, $\mu \sim$ Normal(0,2) and $\sigma \sim$ Half-Normal(0.03), respectively. These results indicate a survival benefit of CRT-D

Table 3.3: CRT-D vs CRT-alone: posterior mean for the overall rate ratio and 95% credible intervals for 8 primary studies under a variety of prior distributions utilizing arm-specific follow-up when available. [a]E(σ) = 0.14; [b]E(σ) = 0.35; [c]E(σ) = 0.41.

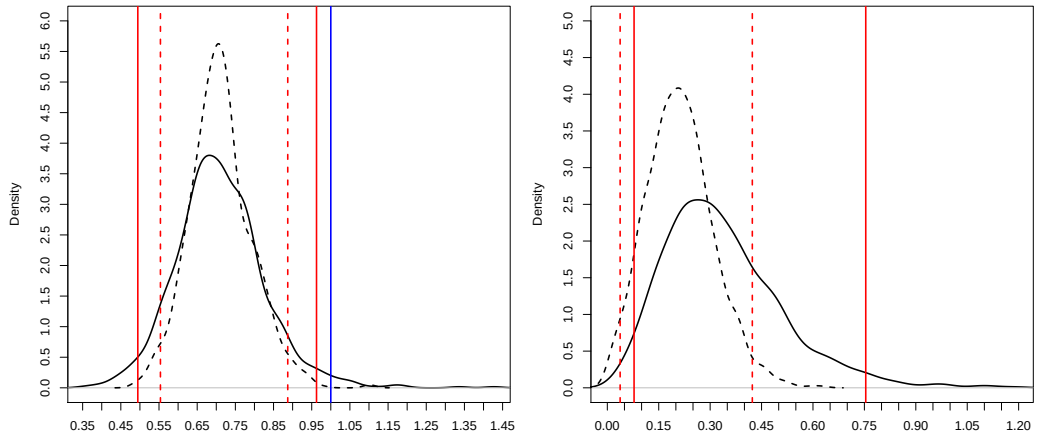
Prior for	Prior for overall underlying log rate ratio, μ		
between-study standard deviation, σ	Normal$(0, 2)$	Normal$(0, 10)$	Normal$(0, 1e06)$
Half-Normal$(0.03)^a$	0.71 (0.55, 0.89)	0.71 (0.56, 0.90)	0.71 (0.55, 0.89)
Uniform$(0, 0.7)^b$	0.71 (0.49, 0.99)	0.71 (0.51, 0.96)	0.72 (0.52, 0.98)
Half-Normal$(0.26)^c$	0.70 (0.50, 0.94)	0.71 (0.51, 0.94)	0.71 (0.49, 0.96)

compared to CRT-alone, such that there is approximately a 30% lower rate of death in the CRT-D arm (Figure 3.3(a)). The posterior mean of the between-study standard deviation was estimated as 0.34 (0.08, 0.75) (Figure 3.3(b)). All other priors resulted in a similar overall rate ratio. The 95% credible interval did not cover 1 for any of the priors (Table 3.3). Using the Half-Normal(0.03) prior for σ resulted in shorter credible intervals (Figure 3.3).

Results are similar, but the overall rate ratio ($\exp(\mu)$) is further from the null when ignoring the arm-specific follow-up information reported in the three studies with a posterior mean of 0.69 (see Appendix for more results). The findings from the simulation studies when $f < 1$ (as in the CRT-D vs CRT meta-analysis) suggest that the estimate ignoring arm-specific information will be further from the null than the estimate using the arm-specific follow-up. It is not surprising that our two posterior means (0.69 and 0.71) do not differ, given only 3 out of 8 studies reported arm-specific follow-up.

3.4 Remarks

We examined the impact of missing duration of follow-up between two treatment arms in meta-analysis on inference about an overall mean. Although events occur at any point over a follow-up period and censoring occurs throughout that period, most applied researchers continue to use odds ratios and assume similar follow-up across treatment groups. Equal follow-up among treatment groups is unlikely to hold in observational



(a) Risk of dying in CRT-D vs. CRT, $\exp \mu$ where thick vertical line indicates null rate ratio of 1.0.

(b) Between study standard deviation, σ.

Figure 3.3: Posterior densities for parameters in the CRT meta-analysis of 8 primary studies. Solid (dashed) lines represent least (most) informative prior distributions for the hyperparameters. Vertical lines represent the 95% credible intervals. Based on 1000 draws from the joint posterior distribution.

studies. In the single study setting, when longer follow-up occurs in the treatment arm, e.g., $f > 1$, the incorrect estimator underestimates the true rate ratio and overestimates the true rate ratio when $f < 1$. Mean squared error is larger relative to the correct estimator when $f \neq 1$ and coverage is poor. Inferences are impacted with a fairly modest difference in follow-up duration, e.g, when $f = 0.8$ and the null is true the coverage of 95% intervals dropped to 0.71 and the bias neared 20%, implying an estimated rate ratio of 0.8.

We utilized hierarchical Poisson regression models to combine rate ratios across studies and examined operating characteristics of posterior means under a variety of conditions using simulation studies. While it is impossible to determine the direction of the bias, both bias and coverage when there is no true effect are worse when using average follow-up – including the available arm-specific information reduces bias compared to including none. However, there is no way to correct the bias unless more information regarding arm-specific follow-up duration is reported. Bushman and Wang (Bushman and Wang (1996)) proposed methods to combine effects when some studies do not report effect estimates using a mixture of the reported effect estimates and vote-counting procedures. Vote-counting procedures require effects to be homogeneous across studies, an assumption not likely to be met in practice. When analyzing the CRT vs. CRT-D studies, there was substantial variability in the duration of follow-up within and between studies, and the majority of the studies did not report duration information. If differences in follow-up between arms exist, then our estimates may over or under-estimate the true rate ratio and are unable to tell in which way.

Moving forward, it is important that publications contain information on arm-specific follow-up duration as applied researchers increasingly combine information from multiple studies to learn about treatment and safety effectiveness. Network meta-analysis may be even more prone to issues of differential follow-up duration because in such analyses the number of treatments and the number of types of studies being compared are large.

Acknowledgements

The authors thank Jennifer Moon, PhD for performing the data abstraction. The authors thank Francesca Dominici, PhD and Miguel Hernan, MD, DrPH for input and suggestions that improved the paper. This material is based upon work supported by the Research Participation Program, Center for Devices and Radiological Health, administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration (FDA), as well as Contract No. HHSF223201110172C and the Critical Path Initiative (Development of Innovative Methodologies for Medical Devices), both from the FDA.

4. A Maximum Likelihood Approach to Power Calculations for the Risk Difference in Stepped Wedge Designs Applied to Left Ventricular Assist Devices

Lauren M. Kunz[1], Sharon-Lise T. Normand[1,2], and Donna Spiegelman[1,3]

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Care Policy, Harvard Medical School
[3]Department of Epidemiology, Harvard School of Public Health

4.1 Introduction

Typically, clinical trials are designed to assess the efficacy of an intervention. After establishing efficacy, effectiveness of the intervention must next be assessed in a large-scale real-life setting. In some instances, the gold standard clinical trials in which individuals are individually randomized may not be feasible or ethical. Cluster randomized trials randomize clusters or groups of people to interventions, rather than individuals. A cluster randomized trial may be more appropriate because of administrative, political, or ethical reasons–for example, there may be a potential for unblinding or contamination of the intervention. However, the cluster randomization can be used to assess efficacy, as well as effectiveness of an intervention.

In this paper, we examine the stepped wedge design (SWD) for cluster randomized trials. The SWD is a type of cluster-level crossover design that begins with no clusters randomized to the intervention and ends with all clusters having the intervention. In the SWD, pre-specified time points are chosen at which clusters are crossed over to the intervention arm. Throughout this paper, the time points will be referred to as steps. Clusters are randomized to the step at which they will receive the intervention, such that all individuals within a cluster at a given step receive the same intervention. Data collection occurs at each step.

The SWD may be of particular utility if it is logistically difficult to implement the intervention simultaneously at many facilities, perhaps due to budget constraints or logistical reasons. SWDs have been implemented across various disciplines (Brown and Lilford (2006)) and may become increasingly used in comparative effectiveness research (CER). These designs permit a rigorous, randomized component for the roll out of a new intervention to clusters, as commonly occurs in practice (Cousens et al. (2011); Mdege et al. (2011); Squire et al. (2011)).

In particular, when studying the comparative effectiveness of medical devices, roll out

or randomization of the time at which a new center receives the newer device makes practical sense. Physicians will need to be trained to use the new devices, as a staggered roll out will permit. For high-risk devices, high-volume centers generally participate in these trials. Within a hospital, physicians will have similar training, work with similar teams of nurses and other hospital staff, and utilize common infrastructure. As such, the outcomes of two patients treated from the same hospital are more likely to be similar than outcomes from two patients treated at different hospitals. The intracluster (or intraclass) correlation (ICC) is defined as the proportion of the total variance of the outcome that is attributable to between-cluster variability, or $\rho = \frac{\tau^2}{\sigma^2+\tau^2}$ where τ^2 is the between-cluster variance and σ^2 is the within-cluster variance. This correlation needs to be taken into account in the design and analysis of a SWD, as in all cluster randomized trials. Consider a study of a left ventricular assist device (LVAD) for patients with end-stage heart disease. As the burden of heart disease increases in the U.S., the number of hearts available for transplant has not increased commensurately and the option of waiting for a transplant versus implantation becomes more relevant, as a heart may never become available. The details of designing a study powered to detect effectiveness of a new type of LVAD will be considered in this paper.

Statistical power is a critical aspect in study design, and is defined as the probability of correctly rejecting the null hypothesis when the alternative hypothesis is true. A literature review by Brown and Lilford (Brown and Lilford (2006)) identified 12 studies implementing a SWD between 1987 and 2005, but only 5 studies reported power/sample size calculations, many of which based power on a cluster randomized design rather than a SWD. In some settings, the cluster randomized designs will over-estimate the power of a SWD, thereby wasting resources at best and prolonging good therapies at worst. If we are to promote the use of the SWD, more research is needed for the statistical properties of the design parameters of these studies. Formulas have been developed for the determination of power in the case of cluster randomized trials with continuous or binary outcomes and are implemented in standard sample size software (Donner and Klar (2000); Hintze

(2008)). Most outcomes in health care trials are binary. However, exact methods that account for the binary nature of outcome data have not been developed to assess power for the SWD. In a two-arm setting, we consider the SWD for binary outcomes aimed at estimating the risk difference as the parameter of interest. We directly compute the variance of the maximum likelihood estimator risk difference to obtain the power of the stepped wedge study design to detect a specified risk difference under different design parameters. We compare our results to the results derived by Hussey and Hughes in their paper on the statistical properties of the SWD (Hussey and Hughes (2007)). Hussey and Hughes develop methodology for continuous outcomes and apply these results to studies with binary outcomes, although the theory underlying their model is not suitable for binary data. We illustrate these results by a SWD for LVAD implantation.

4.2 Methods

4.2.1 The Model

Power calculations for study design depend upon the statistical model assumed to generate the data. We consider a binary treatment and a binary outcome for person k in cluster i at step j, where there are I clusters, J steps for every cluster, and K individuals sampled at each step within a cluster. At each step in a given cluster, new individuals are enrolled, so there are no repeated measurements on individuals. The model assumed for the data is a generalized linear mixed model (GLMM):

$$g(p_{ijk}) = \beta_0 + \beta_1 X_{ijk} + b_i \tag{4.1}$$

where β_0 is the probability of the outcome under standard of care, β_1 is the treatment effect, X_{ijk} is an indicator for the treatment of individual k in cluster i at step j, where $X_{ijk} = X_{ijk'}$ for all i and j, b_i is a random cluster effect, and $E(Y_{ijk} \mid \beta_0, \beta_1, b_i) = p_{ijk}$. Since this is a randomized trial, on average, no confounding is reasonably assumed. There are no time effects, so in the absence of the intervention, the rates of outcome are not

increasing over time. Following Hussey and Hughes, we will focus on the identity link function for $g()$ and a normal distribution for the random effects, $b_i \sim N(0, \tau^2)$.

Intracluster correlation

An important issue in cluster randomized trials, including the SWD, is that outcomes from the same individuals in a cluster are generally positively correlated. This results in an increase in the variance of Y_{ijk} relative to independent data, so the effective sample size is smaller. The intracluster correlation coefficient (ICC), ρ, measures the correlation between individuals from the same cluster. In the above model (4.1), $\rho = \frac{\tau^2}{\tau^2 + p_0(1-p_0)}$ where τ^2 is the variance of the cluster-specific random effects and $p_0(1-p_0) = \sigma_e^2$ is the variance of a Bernoulli random variable ($p_0 = Pr(Y = 1 \mid X = 0) = \beta_0$). Another measure of the cluster effect on the variance is the coefficient of variation, ν, τ/μ for continuous outcomes or τ/p_0 for binary outcomes (Hussey and Hughes (2007)). For binary outcomes, $\frac{\rho}{1-\rho} = \frac{\tau^2}{p_0(1-p_0)} = \nu^2(\frac{p_0}{1-p_0})$. Hence, $\nu = \sqrt{(\frac{\rho}{1-\rho})(\frac{1-p_0}{p_0})}$, $\rho = \frac{\nu^2 p_0}{(\nu^2-1)p_0+1}$, and $\tau^2 = \frac{\rho}{1-\rho}p_0(1-p_0) = \nu^2 p_0^2$. For $p_0 = 0.05$ and a fixed τ, CVs ranging from 0 to 0.5 correspond to ICCs ranging from 0 to 0.013. We generally do not have prior knowledge or a context-relevant estimate of the ICC. Therefore, power is typically is calculated over a range of hypothesized values for the ICC.

4.2.2 Power

The primary object of inference in the SWD is the parameter β_1, the individual level risk difference, and the goal of the study is to test the hypothesis $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 = \beta_A$. We base the power calculations on the Wald test using the asymptotic normal distribution of the maximum likelihood estimator (MLE).

Theoretical power is calculated as

$$\Phi\left(\frac{\beta_A}{\sqrt{Var(\hat{\beta}_1)}} - Z_{1-\alpha/2}\right)$$

where β_A is the value of β_1 under the alternative hypothesis, H_A, $\Phi(.)$ denotes the cumulative normal distribution, and α is the Type I error rate. The challenge is in computing the theoretical variance of $\hat{\beta}_1$, the 2-2 element of $\left(-E\left[\frac{\partial^2 lnL(\theta)}{\partial\theta\partial\theta'}\right]\right)^{-1}$. There are J total steps and K individuals at each step, so for each cluster i there are $J * K = N$ individuals and the total number of individuals for the entire stepped wedge design is $I * J * K$. Because $p_{ijk} = p_{ij'k'}$ due to no time effects, notation can be simplified such that the outcomes for individuals in cluster i are $\mathbf{Y_i} = (Y_{i1}, ..., Y_{iN})$.

Under the linear-normal model, $Pr(Y_{in} = 1) = \beta_0 + \beta_1 X_{in} + b_i$,
$Pr(Y_{in} = 0) = 1 - (\beta_0 + \beta_1 X_{in} + b_i)$, and the full data likelihood is

$$\begin{aligned} L(\boldsymbol{\beta}, \tau^2) &= \prod_{i=1}^{I} \int \prod_{n=1}^{N} f(Y_{in} \mid b_i, \boldsymbol{\beta}) f(b_i) db_i \\ &= \prod_{i=1}^{I} \int \prod_{n=1}^{N} (\beta_0 + \beta_1 X_{in} + b_i)^{Y_{in}} (1 - (\beta_0 + \beta_1 X_{in} + b_i))^{1-Y_{in}} \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{b_i^2}{2\tau^2}} db_i \end{aligned}$$

and the log-likelihood is

$$l(\boldsymbol{\beta}, \tau^2) = \sum_{i=1}^{I} log \int \prod_{n=1}^{N} (\beta_0 + \beta_1 X_{in} + b_i)^{Y_{in}} (1 - (\beta_0 + \beta_1 X_{in} + b_i))^{1-Y_{in}} \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{b_i^2}{2\tau^2}} db_i \quad (4.2)$$

To demonstrate how X_{in} is generated, consider the following example with $J = 3$ steps, $I = 8$ clusters, and $K = 15$ subjects enroll per step. Clusters are randomized to the step at which they receive the intervention and step assignments are fixed before the start of the study. At step 1, $X_{in} = 0$ for all $i = 1, ..., 8$ and $n = 1, ..., 15$. The outcome is measured in the 15 subjects in each cluster, of which none of the $8 \times 15 = 120$ subjects receive the new treatment. At step 2, $X_{in} = 1$ for $i = 1, ..., 4$ and $X_{in} = 0$ for $i = 5, ..., 8$ and $n = 16, ..., 30$, because $I/(J-1) = 4$ clusters are rolled over to the intervention. Now, in 4 of the 8

clusters, all 15 individuals receive the intervention, so $4 \times 15 = 60$ individuals receive the new intervention and 60 do not. At step 3, $X_{in} = 1$ for all $i = 1, ..., 8$ and $n = 31, ..., 45$ and the remaining 4 clusters are rolled over to the intervention, so all 120 individuals having received the intervention. By the end of the study, 180 individuals have received the intervention and 180 have not.

Because there are 4 possible cases for binary X_{in} and Y_{in}: (0,0), (0,1), (1,0), and (1,1), the study data for a single cluster can be reduced to

		outcome, Y_{in}		
		0	1	
intervention, X_{in}	0	Z_{i00}	Z_{i01}	$Z_{i0.}$
	1	Z_{i10}	Z_{i11}	$Z_{i1.}$
				N_i

We rewrite the log-likelihood without the product term over N by utilizing this fact,

$$l(\boldsymbol{\beta}, \tau^2) = \sum_{i=1}^{I} log \int (\frac{1}{\sqrt{2\pi}\tau} e^{-\frac{b^2}{2\tau^2}})^{N_i} \times (1 - (\beta_0 + b))^{Z_{i00}} (\beta_0 + b)^{Z_{i01}} (1 - (\beta_0 + \beta_1 + b))^{Z_{i10}} (\beta_0 + \beta_1 + b)^{Z_{i11}} db_i$$

4.2.3 Theoretical Variance

Asymptotically, as $N = J \times K$ goes to infinity, the variance of $\hat{\beta}_1$ using a maximum likelihood approach is the 2-2 element of $\left(-E\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right]\right)^{-1}$, where $\theta = (\beta_0, \beta_1, \tau^2)$:

$$\begin{aligned} Var(\theta) &\approx [I(\theta)]^{-1} \\ &= (-E[H(\theta)])^{-1} \\ &= \left(-E\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\right]_{3\times 3}\right)^{-1} \qquad (4.3) \\ &= \left(E\left[\left(\frac{\partial l(\theta)}{\partial\theta}\right)_{3\times 1}\left(\frac{\partial l(\theta)}{\partial\theta}\right)^T_{1\times 3}\right]\right)^{-1} \qquad (4.4) \end{aligned}$$

where $H()$ is the Hessian or the matrix of second derivatives of the log-likelihood and $I()$ is the expected Fisher Information (Newey and McFadden (1994)). Since expectation of

the score is 0, equation 4.4 is the variance-covariance matrix of the score equations and by the Cauchy-Schwarz inequality, it will always be positive definite. In greater detail,

$$\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}_{3\times 3} = \left(\sum_i \begin{array}{ccc} \frac{\partial^2 l(\theta)}{\partial\beta_0^2} & \frac{\partial^2 l(\theta)}{\partial\beta_0\partial\beta_1} & \frac{\partial^2 l(\theta)}{\partial\beta_0\partial\tau^2} \\ \frac{\partial^2 l(\theta)}{\partial\beta_1\partial\beta_0} & \frac{\partial^2 l(\theta)}{\partial\beta_1^2} & \frac{\partial^2 l(\theta)}{\partial\beta_1\partial\tau^2} \\ \frac{\partial^2 l(\theta)}{\partial\tau^2\partial\beta_0} & \frac{\partial^2 l(\theta)}{\partial\tau^2\partial\beta_1} & \frac{\partial^2 l(\theta)}{\partial\tau^2{}^2} \end{array} \right) \tag{4.5}$$

and

$$\left(\frac{\partial l(\theta)}{\partial\theta}\right)_{3\times 1}\left(\frac{\partial l(\theta)}{\partial\theta}\right)^T_{1\times 3} = \left(\begin{array}{c} \frac{\partial l(\theta)}{\partial\beta_0} \\ \frac{\partial l(\theta)}{\partial\beta_1} \\ \frac{\partial l(\theta)}{\partial\tau^2} \end{array} \right)\left(\begin{array}{c} \frac{\partial l(\theta)}{\partial\beta_0} \\ \frac{\partial l(\theta)}{\partial\beta_1} \\ \frac{\partial l(\theta)}{\partial\tau^2} \end{array} \right)^T = \left(\sum_i \begin{array}{ccc} (\frac{\partial l(\theta)}{\partial\beta_0})^2 & \frac{\partial l(\theta)}{\partial\beta_0}\frac{\partial l(\theta)}{\partial\beta_1} & \frac{\partial l(\theta)}{\partial\beta_0}\frac{\partial l(\theta)}{\partial\tau^2} \\ \frac{\partial l(\theta)}{\partial\beta_1}\frac{\partial l(\theta)}{\partial\beta_0} & (\frac{\partial l(\theta)}{\partial\beta_1})^2 & \frac{\partial l(\theta)}{\partial\beta_1}\frac{\partial l(\theta)}{\partial\tau^2} \\ \frac{\partial l(\theta)}{\partial\tau^2}\frac{\partial l(\theta)}{\partial\beta_0} & \frac{\partial l(\theta)}{\partial\tau^2}\frac{\partial l(\theta)}{\partial\beta_1} & (\frac{\partial l(\theta)}{\partial\tau^2})^2 \end{array} \right) \tag{4.6}$$

and the sums are over i, the clusters, where the clusters are independent.

Let $\mathbf{Z} = (Z_{00}, Z_{01}, Z_{10}, Z_{11})$, which is the only random component after integrating out the unobserved random cluster effect. Given β_0, β_1, τ^2 through knowledge of the expected baseline rate, a desired effect size and the intracluster correlation, and given the number of clusters I, steps J and sample sizes N per cluster. We find the expected value of the elements of equations (4.5) and (4.6). Because both the intervention and outcome are binary, we sum over all $\mathbf{Z}$, for example $E[\frac{\partial^2 lnL_i}{\partial\theta^2}(\mathbf{Z})] = \sum_{Z_{10}=0}^{Z_{1\cdot}}\sum_{Z_{00}=0}^{Z_{0\cdot}}\frac{\partial^2 lnL_i}{\partial\theta^2}(\mathbf{Z})Pr(\mathbf{Z})$ and $E[(\frac{\partial lnL_i}{\partial\theta}(\mathbf{Z}))^2] = \sum_{Z_{10}=0}^{Z_{1\cdot}}\sum_{Z_{00}=0}^{Z_{0\cdot}}(\frac{\partial lnL_i}{\partial\theta}(\mathbf{Z}))^2 Pr(\mathbf{Z})$ where $Z_{1\cdot}$ is the number of people in cluster i that receive the intervention and $Z_{0\cdot}$ is the number of people in cluster i that are untreated, both of which are fixed by design. Further, $Z_{1\cdot} = \frac{j_i}{J}N, Z_{0\cdot} = (1-\frac{j_i}{J})N$ where j_i is the number of steps in cluster i on treatment. Hence, if Z_{10} varies from 0 to $Z_{1\cdot}$, Z_{11} will simultaneously be varying from $Z_{1\cdot}$ to 0, since $Z_{1\cdot} = Z_{11} + Z_{10}$.

Using properties of general matrix inversion to obtain the 2-2 element,

$$Var(\hat{\beta}_1) = \frac{h_{33}h_{11} - h_{13}^2}{h_{11}h_{22}h_{33} + 2 * h_{12}h_{23}h_{13} - h_{13}^2h_{22} - h_{12}^2h_{33} - h_{23}^2h_{11}} \tag{4.7}$$

where the denominator is the determinant of a 3×3 matrix and h_{rc} is the expected value of the rth row and cth column of equations (4.5) and (4.6) with expectations taken over $\mathbf{Z} \mid Z_{1.}, Z_{0.}$.

Four unique integrals are required to calculate the elements that comprise the outer product of first derivatives of the log-likelihood (an additional 6 are required for the matrix of second derivatives). The details of these complex integrals can be found in the appendix. For each unique treatment pattern of the SWD, where there are $J-1$ unique intervention patterns and $I/(J-1)$ clusters randomized to each pattern, we find the 2-2 element of equation (4.4) as follows:

1. Calculate all integrals for each possible realization of $\mathbf{Z}$ over b
2. Combine these into the first derivatives following the equations in the appendix to obtain $\frac{\partial l_i}{\partial \theta}(\mathbf{Z})$
3. Multiply the first derivatives by

$$Pr(\mathbf{Z}) = \frac{\int (\frac{1}{\sqrt{2\pi\tau}} e^{-\frac{b^2}{2\tau^2}})^N \times (1-(\beta_0+b))^{Z_{00}} (\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}} db}{\sum\limits_{Z} \int (\frac{1}{\sqrt{2\pi\tau}} e^{-\frac{b^2}{2\tau^2}})^N \times (1-(\beta_0+b))^{Z_{00}} (\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}} db}$$

4. Sum over all realizations of $\mathbf{Z}$ to obtain $\sum\limits_{Z_{10}=0}^{Z_{1.}} \sum\limits_{Z_{00}=0}^{Z_{0.}} (\frac{\partial l_i}{\partial \theta}(\mathbf{Z}))(\frac{\partial l_i}{\partial \theta}(\mathbf{Z}))Pr(\mathbf{Z})$
5. Sum over all clusters i to obtain$\sum\limits_i E[(\frac{\partial l_i}{\partial \theta}(\mathbf{Z}))^2]$ (note that for independent clusters the sum of the expectations over i is equal to the expectation of the sums of the derivatives over i), which are the h_{rc} elements
6. Calculate $Var(\hat{\beta}_1)$ as equation (4.7)

Increasing either N or J increases computation time as we must compute the integrals in step 1 for each intervention pattern and the resulting possible realizations of

$\mathbf{Z} = (Z_{0\cdot} + 1)(Z_{1\cdot} + 1)$ which depend on N. The computing time for this method is high to achieve convergence at double precision accuracy using Romberg integration (see appendix section A.3.2 for the computational details).

4.2.4 Hussey and Hughes method

Our model (equation 4.1) is the same as that assumed by Hussey and Hughes, except they assume a continuous outcome Y_{ijk} and have $J-1$ additional fixed effects for a time effect at each step. Given τ^2 and σ^2, their variance formula is based on a weighted least squares estimator with weights based on an exchangeable within-cluster correlation structure. This gives a closed form variance estimator:

$$Var(\hat{\beta}_1) = \frac{I\sigma^2(\sigma^2 + J\tau^2)}{(IU - W)\sigma^2 + (U^2 + IJU - JW - IV)\tau^2} \tag{4.8}$$

where $\sigma^2 = \frac{p_0(1-p_0)}{K}$, $U = \sum_{ij} X_{ij}$, $W = \sum_j (\sum_i X_{ij})^2$ and $V = \sum_i (\sum_j X_{ij})^2$ where $X_{ij} = 1$ if cluster i receives the intervention at step j.

When no time effects are assumed, as here, equation (4.8) reduces to

$$Var(\hat{\beta}_1) = \frac{I\frac{\sigma_e^2}{K}(\frac{\sigma_e^2}{K} + J\tau^2)}{(IJU - U^2)\frac{\sigma_e^2}{K} + IJ(JU - V)\tau^2}$$

When $Y_{ijk} \sim Bernoulli(p_{ijk})$, Hussey and Hughes assume $Var(Y_{ijk}) = \sigma_e^2 + \tau^2$, $Var(Y_{ij}) = \frac{\sigma_e^2}{K} + \tau^2 = \sigma^2 + \tau^2$, and $\sigma^2 = \frac{p_0(1-p_0)}{K}$ (K is the number of individuals sampled at each step within a cluster).

4.3 Design parameters & Results

To explore the properties of our method proposed in section 4.2.3, we examined the behavior of Var($\hat{\beta}_1$) as effect sizes ranged from large to small, 0.1 to 0.0125 (β_1) with $\beta_0 = 0.05$

to give a variety of effects on the risk difference scale. We considered two intracluster correlation coefficients (ICCs), 0.001 and 0.01, to represent a small and a moderate correlation for binary outcomes. The total number of clusters (I) ranged from 8 to 80 and two step wedge randomization patterns were considered (J=3 and 5). Cluster sizes of $N = 15, 45,$ and 90 were considered. From section 4.2.1, once β_0 and the ICC are specified, τ^2 can be calculated. The integrals in the appendix can then be calculated for every realization of $\mathbf{Z}$ for the given design and summed over all clusters i to obtain the variance of the estimated risk difference, β_1 under a given design, and then, the power. The parameter choices above led to 32 different designs in total. We assessed compared power by examining the asymptotic relative efficiency defined as the ratio of the variance of the risk difference parameter for the Hussey and Hughes method to our maximum likelihood method: $\frac{Var(\widehat{\beta_{1,HH}})}{Var(\widehat{\beta_{1,ML}})}$. This can be interpreted as the factor by which the sample size would need to be increased if HH was used for design instead of our ML method.

4.3.1 Comparison to Hussey and Hughes (HH)

Over the range of design parameters considered, for a fixed number of patients, our maximum likelihood (ML) method provided designs that ranged from 9% to 2.4 times more efficient than designs based on HH as measured by $\frac{Var(\widehat{\beta_{1,HH}})}{Var(\widehat{\beta_{1,ML}})}$.

The ARE increases as the effect size increases, as the ICC increases, for an increased number of steps (J), and as the number of patients per cluster (N) increases.

4.3.2 General observations

As expected, decreasing the magnitude of the effect size to consider risk differences from 0.1 to 0.0125, resulted in decreased power. There was a larger decrease in power when the total number of clusters was smaller when decreasing the effect size relative to decreasing the effect size for a larger sample size (having more total clusters) (see Figure 4.1).

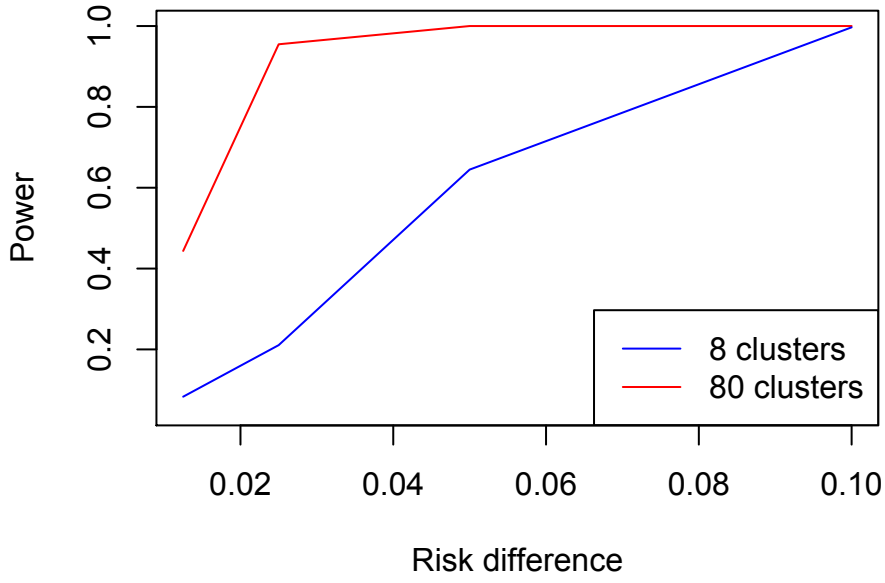


Figure 4.1: Power in relation to the effect size, with a baseline risk of 0.05, 90 individuals per cluster, 3 steps, and an ICC=0.01. For I=8 clusters, the total sample size is 720 and for I=80, the total sample size is 7200.

Table 4.1: Asymptotic relative efficiency (ARE)= $\frac{Var(\widehat{\beta_{1,HH}})}{Var(\widehat{\beta_{1,ML}})}$ comparing the SWD to HH, with a baseline risk of 0.05 and I=8 total clusters. RD=risk difference, ICC=intracluster correlation coefficient, J=number of steps, N=total sample size per cluster over all steps

Risk Difference	ICC	J Steps	N Patients / Cluster	ARE
0.05	0.001	3	90	1.77
0.05	0.01	3	90	2.05
0.05	0.001	5	90	2.07
0.05	0.01	5	90	2.47
0.0125	0.01	3	90	1.99
0.025	0.01	3	90	2.01
0.05	0.01	3	90	2.05
0.10	0.01	3	90	2.07
0.05	0.01	3	45	1.09
0.05	0.01	3	90	2.05
0.05	0.01	5	45	1.25
0.05	0.01	5	90	2.47

We replicated many of the trends reported by Hussey and Hughes, but with increased power for our maximum likelihood method. Hussey and Hughes showed that power was relatively insensitive to varying the ICC, parametrized by the coefficient of variation, with slightly larger power for smaller ICCs. The difference in the variance of β_1 was less than 0.01% across the ICCs considered (results not shown), which did not impact power.

If fewer clusters are crossed over to the intervention at each step, Hussey and Hughes noted that power increases (Hussey and Hughes (2007)). For fixed total sample size, there is a trade off between the number of steps and number of clusters switched to the intervention at each step. Power increased when the number of steps increased (see figure 4.2). The increase in power by increasing the number of steps also occurred at a risk difference of 0.025 and an ICC of 0.001 (results not shown).

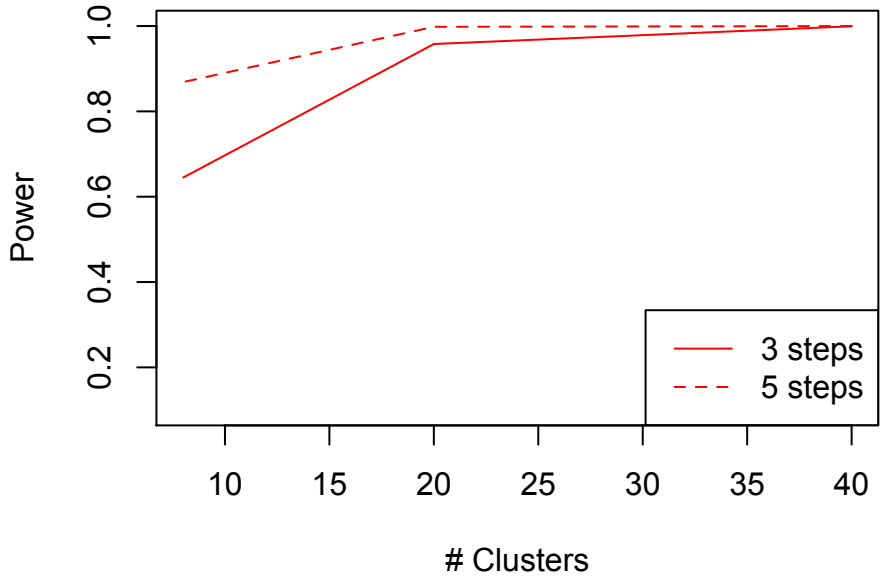


Figure 4.2: Power in relation to the number of steps (J), at fixed $N = 90$ individuals per cluster, with a baseline risk of 0.05, risk difference of 0.05, ICC=0.01. As the number of clusters increases, so does the total sample size.

4.3.3 Comparison to general cluster randomized design (CRD)

Additionally, it is also of interest to compare the SWD to a cluster randomized trial (Donner and Klar (2000)). Fixing the total number of clusters and assuming half were randomized to the intervention and half to the control all at the same time point, we compared the power of the SWD to the CRD for an equal number of patients receiving the intervention, number of clusters, and ICC.

When the number of patients per cluster was small (N=15), the CRD had higher power than the equivalent SWD. At a moderate number of patients per cluster (N=45), the number of steps determined whether the CRD or the SWD was more powerful, suggesting that for a given design there will be a point at which by increasing the number of steps, the SWD will be more powerful, at the possible expense of extending the length of the study (see table 4.2). We also note that the CRD is strongly sensitive to the ICC.

Table 4.2: Power for the SWD versus CRD with a baseline risk of 0.05. Assume both designs have the same total number of clusters and total sample size. RD=risk difference, ICC=intracluster correlation coefficient, I=number of clusters, J=number of steps, N=total sample size per cluster over all steps

					Power	
Risk Difference	ICC	I Clusters	J Steps	N Patients / Cluster	SWD	CRD
0.05	0.01	8	3	15	0.086	0.165
0.05	0.01	80	3	15	0.471	0.872
0.05	0.01	8	3	45	**0.235**	**0.326**
0.05	0.01	80	3	45	0.975	0.998
0.05	0.01	8	5	45	**0.362**	**0.326**
0.05	0.01	80	5	45	0.999	0.998
0.05	0.001	8	5	90	0.869	0.689
0.05	0.01	8	5	90	0.869	0.461
0.1	0.01	8	3	90	0.997	0.910
0.05	0.01	8	3	90	0.645	0.461
0.025	0.01	8	3	90	0.211	0.172
0.0125	0.01	8	3	90	0.084	0.083

4.4 Example: LVAD study design

Left ventricular assist devices (LVADs) are implantable devices that may be used as a bridge for patients who are awaiting heart transplants or a destination for those who are not eligible for a transplant. In either case, these patients suffer from severe heart disease that cannot be controlled with any other therapy. LVADs are life saving devices, so both doctors and patients may be reluctant to participate in standard randomized trial designs, as withholding implantation of an LVAD can be deemed unethical. A 2006 working group recommended that designs for LVADs should include a randomized component (Neaton et al. (2007)). The SWD is a feasible option for evaluating a new LVAD. Furthermore, there are likely large center effects for this very complex procedure that requires highly skilled surgeons. As of 2010, only 69 medical centers were certified for implementation of LVADs as destination therapy by the Centers for Medicare and Medicaid (Kirklin et al. (2011)).

The Randomized Evaluation of Mechanical Assistance for the Treatment of Congestive Heart Failure (REMATCH) trial showed that an LVAD was superior to medical therapy as destination therapy (Rose et al. (2001)). In the LVAD group, there were 41 deaths among 68 patients (60.3%) implanted, and in the medical therapy group there were 54 deaths out of 61 patients (88.5%) over 20 experienced centers. Based on this trial, we are interested in the design of hospital randomized trial to implant LVADs as destination therapy, where all patients eligible for an LVAD would be implanted in a given hospital or none of the patients at that hospital could be implanted once the hospital reaches the step at which it was randomized to receive the LVAD. The primary endpoint is 6-month survival and the parameter of interest is the risk difference. We assume a baseline proportion surviving of 20%. Another destination therapy trial compared the pulsatile Heartmate XVE with the continuous-flow Heartmate II (Slaughter et al. (2009)). Use of LVADs for destination therapy has increased tenfold the previous amount since January, 2010 when Heartmate II for destination therapy was approved (Stewart and Stevenson (2011)). At the Mayo Clinic, 117 patients underwent LVAD implementation as destination therapy from Febru-

ary 2007 to June 2012 (64 months) (Dunlay et al. (2014)). In addition, according to the Interagency for Mechanically Assisted Circulatory Support (INTERMACS) registry, for destination therapy, 135 LVADs were placed from 2006 to 2009, 464 in 2010 and 248 in the first 6 months of 2011 (Kirklin et al. (2012)). Based on these numbers, it is reasonable to assume that these high volume centers could actively each identify 1-3 viable candidates for LVAD destination therapy per month to accrue into a stepped wedge study design. We assume that there are a number of medical centers awaiting approval to implant LVADs as destination therapy with appropriate teams in place. We will compute the power for randomizing every 6 months over a period of 3 years, and examine the impact of adding additional medical centers. This corresponds to 7 steps in the SWD with no centers performing destination therapy at baseline and at 3 years all centers will implant LVADs as destination therapy for all eligible patients.

Power analyses deal with many uncertainties, including the ICC, the values of the cluster sizes, and the effect size. For this example, no data from previous studies reported ICCs. Because the SWD is insensitive over a range of reasonable values of the ICC, we consider an ICC of 0.01 for this application. We assume that there are 12 medical centers that will participate in this SWD and over each 6 month step 10 patients would meet the criteria to be eligible for destination therapy, such that each medical center has a total sample size of 70 patients across the entire three year period, leading to a total sample size of 840 for the entire SWD.

For a baseline 6 month survival of 20% in the control group, the variance of the risk difference was 1.786×10^{-4}, which will result in 96% power to detect a risk difference of 0.05 or greater, indicating at least 5% greater 6 month survival for those implanted with LVAD for destination therapy. Hussey and Hughes method yielded 93% power for this effect size. A CRD with 6 clusters randomized to each intervention with 70 patients per cluster would only have 27% power to detect a risk difference of 5% (calculated using PASS Hintze (2008)).

4.5 Discussion

In this paper, we computed theoretical asymptotic power for a binary outcome in a SWD. This involved difficult integrations over the distribution of the unobservable random cluster effects. By doing so, we were able to appropriately account for binary nature of the outcome data using maximum likelihood theory, and showed that under several different study designs, the resulting power was greater than that found by Hussey and Hughes, using their closed form approximation for binary data. Over the range of design parameters considered, for a fixed number of patients, our exact maximum likelihood method provided designs that ranged from 9% to 2.4 times more efficient than designs based on Hussey and Hughes' variance approximation. We note that Hussey and Hughes does have the advantage of being a closed form solution. However, with current computational capabilities our method, which will be made freely available, results in more efficient designs.

We also found that for the SWD, power is insensitive to variations in assumptions about the intracluster correlation coefficient. For a fixed overall number of clusters and individuals per cluster, increasing the number of steps leads to increases in the power. When we compared the SWD to the cluster randomized design, there was a point at which increasing the number of steps led to the SWD becoming more powerful than a CRD with the same number of subjects, clusters, ICC and marginal treatment proportion. The SWD may be longer and potentially more expensive than the CRD, however. The utility of the SWD may be in the ability to add a randomized component when it is unethical to withhold the intervention from patients, when the argument may be that any randomization is better than none when causal effect estimates are the goal.

As mentioned, little statistical theory for stepped wedge study design exists and there are many extensions. We would like to extend our method to include settings with a time effect. In many areas of application, the inclusion of time effects will be important. In our LVAD example, if eligibility requirements are restructured this could lead to better

outcomes over time in both the LVAD and optimal medical therapy groups. Time by treatment interactions may also need to be taken into account, for example, in our LVAD example, there is likely a learning curve effect where surgeons will take some time to become more familiar and comfortable with the use of the device or the procedure, perhaps leading to improvements in the outcome.

We used the identity link to specify the effect on the risk difference scale. Extensions to the log and logistic links, when the parameter of interest is the risk ratio and odds ratio will be required. We realize that the normal distribution we have placed on the probability has infinite support which is not realistic for a probability, but it is likely that because the variance is small, when we evaluate the integral by Romberg quadrature, the range of the integral is sufficiently large to incorporate most of the mass and at the same time retain the constraint that the probability is restricted to range between zero and one. Further research will evaluate the impact of this approximation on the SWD variance calculations and subsequent power and sample size outputs. A constrained likelihood can be derived based upon a truncated mean zero normal which retains the flavor of normal random effects as is commonly assumed but satisfies the restrictions relevant to binomial data. In addition, simulation studies will be conducted that compare the empirical variance of the constrained and unconstrained models to assess the impact of the approximation.

Simulation based methods are another approach for power calculations. We opted for an exact approach using Romberg integration. For the same precision, we would like to compare the computation efficiency of the two approaches and expect to find that the exact method would be considerably faster than simulation based methods for a fixed amount of precision.

In summary, we have demonstrated that the asymptotic maximum likelihood approach to power calculations provides more efficient study designs for detecting a risk difference of pre-specified magnitude than the previously available method. This suggests that the SWD may be a more feasible study design than has been previously appreciated. We have

focused on design components, but our findings also have implications for the analysis as well, suggesting that the maximum likelihood estimator will often be substantially more efficient than GEE for estimating the risk differences of SWDs.

Acknowledgements

Dr. Spiegelman and Ms. Kunz were supported by grants from the National Institutes of Health (1R01AI112339 and 1DP1ES025459). Dr. Normand's effort was supported by a grant from the US Food and Drug Administration U01FD004493.

A. Appendices

A.1 An Overview of Statistical Approaches for Comparative Effectiveness Research for Assessing In-Hospital Complications of Percutaneous Coronary Interventions By Access Site

A.1.1 Factors associated with Radial Artery Access vs Femoral Artery Access

Table A.1: Covariates included in the propensity score model.

Linear Terms	Interaction Terms
Female	Smoker:Race
Diabetes	Smoker:Age
Smoker	Smoker:Platelet Aggregate Inhibitors
Prior PCI	Prior CABG:Peripheral Vascular Disease
Prior MI	Prior CHF:Hypertension
Prior CABG	Prior CHF:G2B3A Inhibitors
Prior CHF	Prior CHF:Thrombin
Lung Disease	Lung Disease:Left Main Disease
STEMI	Lung Disease:Hypertension
Race: white (baseline), black, hispanic, other	Age^2
Insurance: government (baseline), commercial, other	STEMI:Fractionated Heparin
Shock	STEMI:Low Molecular Weight Heparin
Left Main Disease	STEMI:G2B3A Inhibitors
Age	STEMI:Platelet Aggregate Inhibitors
Multi-vessel Disease	STEMI:Thrombin
Number of Vessels $>$ 70% stenosis	Insurance:Race
Peripheral Vascular Disease	Insurance:Age
Hypertension	Insurance:Peripheral Vascular Disease
Aspirin	Insurance:Hypertension
Fractionated Heparin	Insurance:G2B3A Inhibitors
Low Molecular Weight Heparin	Insurance:Thrombin
G2B3A Inhibitors	Age:Peripheral Vascular Disease
Platelet Aggregate Inhibitors	Age:Hypertension
Thrombin	Age:Platelet Aggregate Inhibitors
	Age:Thrombin
	Age:Fractionated Heparin
	Age:Low Molecular Weight Heparin
	Peripheral Vascular Disease:Fractionated Heparin
	Peripheral Vascular Disease:Low Molecular Weight Heparin
	Fractionated Heparin:G2B3A Inhibitors
	Low Molecular Weight Heparin:G2B3A Inhibitors
	Fractionated Heparin:Platelet Aggregate Inhibitors
	Low Molecular Weight Heparin:Platelet Aggregate Inhibitors
	Fractionated Heparin:Thrombin
	Low Molecular Weight Heparin:Thrombin
	G2B3A Inhibitors:Thrombin

A.1.2 R code

All analyses were performed with R software, version 2.14.1. General code is provided for the case where Y is the binary outcome, T is the binary treatment, X is a vector of covariates, and PS is the estimated propensity score. All standard errors were computed by bootstrapping. The appendix of Ahern et al. (Ahern et al. (2009)) provides guidance for this procedure.

Propensity Score Estimation. Denote the propensity score by PS and the linear propensity score by lPS.

```
PSmodel=glm(T ~ X,family=binomial(link="logit"),data=dataset)
PS=predict(PSmodel,dataset,type="response")

lPS=predict(PSmodel,dataset)
```

Matching on the Propensity Score. Using the Matching package to perform 1-1 matching (M=1), without replacement (replace=FALSE) and a caliper of 0.2 standard deviations of the propensity score (caliper=0.2), to estimate the average treatment effect (estimand="ATE"). Further options can be found in the manual for this package.

```
library(Matching)
runmatch=Match(Y=Y,Tr=T,X=lPS, M=1,replace=FALSE,caliper=0.2,estimand="ATE")
runmatch$est # estimated ATE
runmatch$est-1.96*runmatch$se.standard # lower 95% CI limit

runmatch$est+1.96*runmatch$se.standard # upper 95% CI limit
```

The original data can be accessed to identify the matched pairs using:

```
matcheddata = dataset[c(runmatch$index.treated,runmatch$index.control),]
```

Stratification on the Propensity Score. First create the quintiles and then create a variable to indicate the stratum to which a subject belongs. The data may be divided into fewer or

more quantiles by modifying the quantile() command. Balance within each stratum can be assessed with a t-test, where in the following "i" should be replaced by the stratum of interest. A loop can be used to quickly cycle through all of the strata.

```
breakvals=quantile(PS, prob=0:5*0.2)
strat=cut(PS, breaks=breakvals, labels=c('1','2','3','4','5'),include.lowest=TRUE)

t.test(PS[strat==i&T==1],PS[strat==i&T==0])
```

To combine the results across strata, we wrote the following functions that can be called by plugging in the variables for the outcome (out), treatment (treat), and strata (str).

```
difference.means = function(out, treat)

{mean(out[treat==1], na.rm=TRUE) - mean(out[treat==0], na.rm=TRUE)}
```

```
SE = function(out,treat)

{sqrt(var(out[treat==1], na.rm=TRUE)/sum(treat==1)+ var(out[treat==0], na.rm=TRUE)/sum(treat==0))}
```

```
strata.average = function(out, treat, str) {
Q = length(table(str)); n=length(out); differences=rep(NA,Q)
for (q in 1:Q) differences[q] = difference.means(out[str==q],treat[str==q])
weights=table(str)/n
overall.difference = weights%*%differences

return(list("Mean Difference within Strata"=differences,"Average Weighted Mean

Difference"=overall.difference))}
```

```
strata.variance = function(out,treat,str)
{Q = length(table(str)); n=length(out); variances=rep(NA,Q)
for (q in 1:Q) variances[q]= SE(out[str==q],out[str==q])**2
weights = table(str)/n
overall.variance = weights**2%*%variances
return(list("Variance within Strata"=variances,"Overall Variance"=

overall.variance,"Overall SE"=sqrt(overall.variance)))}
```

```
strata.average(Y,T,strat)
strata.variance.average(Y,T,strat)
```

Weighting by the Propensity Score. For the IPTW estimators, the point estimates are obtained using

```
HT.IPTW = mean((T/PS-(1-T)/(1-PS))*Y)
S.IPTW= sum(T*Y/PS)/sum(T/PS)-sum((1-T)*Y/(1-PS))/sum((1-T)/(1-PS))
```

G-computation. G-computation performs the outcome regression and then predicts the outcome as if all subjects received treatment and also if all subjects received control. Without using MSM, the estimate of the ATE is the average of each individuals predicted Y_{1i} and Y_{0i}.

```
outreg=lm(Y ~ T+X, data=dataset)
all.pred=predict(outreg) # predicts Y
T1.pred =predict(outreg,newdata=data.frame(dataset[,-"T"],T=1)) # predictions when T=1 for all subjects
T0.pred = predict(outreg,newdata=data.frame(dataset[,-"T"],T=0)) # predictions when T=0 for all subjects
G.comp = mean(T1.pred-T0.pred)
```

Augmented-IPTW. Augmented IPTW uses the same predictions of the outcome as G-computation and then combines the predictions to compute the point estimate.

```
A.IPTW=mean((T/PS-(1-T)/(1-PS))*(Y-all.pred))+mean(T1.pred-T0.pred)
```

Targeted Maximum Likelihood Estimation. TMLE uses the tmle program and we demonstrate how to specify the parametric forms of the outcome and treatment regressions. Super learning is the default when Qform and gform are unspecified.

```
library(tmle)
TMLE =tmle(Y=Y,A=T,W=X,Qform=Y∼A+X,gform=A ∼X)

summary(TMLE)
```

A.2 Comparative Effectiveness and Meta-Analysis of Cardiac Resynchronization Therapy Devices: The Role of Differential Follow-up

A.2.1 CRT Data: Detailed Follow-up

Table A.2: CRT-D versus CRT-alone studies: Detailed follow-up information reported in studies. Q1 and Q3 are the first and third quartiles, respectively. The ratio of follow-up by treatment arm is denoted $f = \bar{e_1}/\bar{e_0}$.

	Months of Follow-up (unless specified otherwise)			
Study	Overall Average	CRT-D Arm	CRT Arm	$f = \frac{\overline{e_1}}{\overline{e_0}}$
Adlbrecht et al. (2009)	mean 16.8 $\pm$ 12.4			
Stabile et al. (2009)	mean 58 $\pm$ 15 (Q1 49 and Q3 67)	median 56.8	median 60.1	0.95
Bai et al. (2008)	mean 811.6 days $\pm$ 536.7 (range 371 to 2427)			
Auricchio et al. (2007)	median 34 (Q1 10 and Q3 40)			
Ermis et al. (2004)	mean 13.5 $\pm$ 12.0	13 $\pm$ 11.8 (range 4 to 60)	18 $\pm$ 13.2 (range 0.5 to 53)	0.72
Pappone et al. (2003)	mean 840 $\pm$ 257 days			
Bristow et al. (2004)		median 16.0	16.5	0.97
Schuchert et al. (2013)	12			

A.2.2 Bias of the Single Study Estimator for the Rate Ratio

Assume the number of events from treatment arm j is s $Y_j \sim \text{Pois}(\theta_j)$ where $\theta_j = \lambda_j \times \bar{e}_j \times n_j$ with $\bar{e}_j$ the average follow-up in months in arm j and $n_j =$ is the number of subjects in arm j. Then the mortality rate is written as

$$\lambda_j = \xi \exp(\omega \times j) \tag{4.1}$$

where $j = 0$ for the control arm and $j = 1$ for the treatment arm. The maximum likelihood estimator for ω is $\log(\hat{\lambda_1}/\hat{\lambda_0}) = \log\left(\frac{Y_1/\overline{e_1}n_1}{Y_0/\overline{e_0}n_0}\right)$ because $\hat{\lambda_j} = \frac{Y_j}{\overline{e_j}n_j}$. When average follow-up

is the same in each treatment arm, then $\omega = \log\left(\frac{Y_1/n_1}{Y_0/n_0}\right)$. The rate ratio is $\exp(\omega) = \lambda_1/\lambda_0$.

We are dealing with the expectation and variance of ratios of random variables and by Taylor series expansions, it can be shown that

$$\begin{aligned} \mathrm{E}\left(\frac{Y_1}{Y_0}\right) &\approx \frac{EY_1}{EY_0} - \frac{\mathrm{Cov}(Y_1, Y_0)}{E^2Y_0} + \frac{\mathrm{Var}(Y_0)EY_1}{E^3Y_0} \qquad (4.2) \\ \mathrm{Var}\left(\frac{Y_1}{Y_0}\right) &\approx \frac{1}{E^2Y_0}Var(Y_1) + 2\frac{-EY_1}{E^3Y_0}\mathrm{Cov}(Y_1, Y_0) + \frac{E^2Y_1}{E^4Y_0}\mathrm{Var}Y_0 \\ &= \frac{E^2Y_1}{E^2Y_0}\left[\frac{\mathrm{Var}(Y_1)}{E^2Y_1} - 2\frac{\mathrm{Cov}(Y_1, Y_0)}{EY_0\ EY_1} + \frac{\mathrm{Var}(Y_0)}{E^2Y_0}\right] \qquad (4.3) \end{aligned}$$

when expanding out to 3 terms.

Noting that the moments of the Poisson distribution are $E(Y_j) = \theta_j$ and $E(Y_j^2) = \theta_j + \theta_j^2$, then $\mathrm{Var}(Y_j) = \theta_j$ and $E(Y_j^3) = \theta_j^3 + 3\theta_j^2 + \theta_j$. Because Y_1 and Y_0 are independent, $\mathrm{Cov}(Y_1, Y_0) = 0$. When average follow-up by treatment arm, $\overline{e_j}$, is available, then $E(\hat{\lambda_j}) = E(\frac{Y_j}{\overline{e_j}n_j}) = \frac{1}{\overline{e_j}n_j}E(Y_j) = \frac{\theta_j}{\overline{e_j}n_j}$.

Under the model,

$$E(\hat{\lambda_1}) = \frac{\xi\exp(\omega)\overline{e_1}n_1}{\overline{e_1}n_1} = \xi\exp(\omega) \text{ and } E(\hat{\lambda_0}) = \frac{\xi\overline{e_0}n_0}{\overline{e_0}n_0} = \xi$$

Hence,

$$\begin{aligned} E(\widehat{\exp(\omega)}) &= E\left(\frac{\hat{\lambda_1}}{\hat{\lambda_0}}\right) = \frac{\overline{e_0}n_0}{\overline{e_1}n_1}E\left(\frac{Y_1}{Y_0}\right) \\ &\approx \frac{\overline{e_0}n_0}{\overline{e_1}n_1}\left(\frac{\theta_1}{\theta_0} - 0 + \frac{\theta_0 * \theta_1}{\theta_0^3 + 3\theta_0^2 + \theta_0}\right) \\ &= \frac{\overline{e_0}n_0}{\overline{e_1}n_1}\left(\frac{\xi\exp(\omega)\overline{e_1}n_1}{\xi\overline{e_0}n_0} - 0 + \frac{\xi\overline{e_0}n_0 * \xi\exp(\omega)\overline{e_1}n_1}{(\xi\overline{e_0}n_0)^3 + 3(\xi\overline{e_0}n_0)^2 + (\xi\overline{e_0}n_0)}\right) \\ &= \exp(\omega) + \overline{e_0}n_0\left(\frac{\xi\exp(\omega)}{(\xi\overline{e_0}n_0)^2 + 3(\xi\overline{e_0}n_0) + 1}\right) \\ &= \exp(\omega)\left[1 + \frac{\xi\overline{e_0}n_0}{(\xi\overline{e_0}n_0)^2 + 3(\xi\overline{e_0}n_0) + 1}\right]. \qquad (4.4) \end{aligned}$$

Therefore the bias, $E(\widehat{\exp(\omega)}) - \exp(\omega)$ is $\frac{\xi\overline{e_0}n_0\exp(\omega)}{(\xi\overline{e_0}n_0)^2+3(\xi\overline{e_0}n_0)+1}$.

In the absence of arm-specific follow-up, and only average follow-up for the study, $E(\hat{\lambda_j^*}) = E(\frac{Y_j}{\overline{e}n_j}) = \frac{1}{\overline{e}n_j}E(Y_j) = \frac{\theta_j}{\overline{e}n_j}$ and

$$E(\hat{\lambda_1^*}) = \frac{\xi \exp(\omega)\overline{e_1}n_1}{\overline{e}n_1} = \frac{\xi \exp(\omega)\overline{e_1}}{\overline{e}} \text{ and } E(\hat{\lambda_0^*}) = \frac{\xi\overline{e_0}n_0}{\overline{e}n_0} = \frac{\xi\overline{e_0}}{\overline{e}}.$$

Hence,

$$\begin{aligned}
E(\widehat{\exp(\omega^*)}) &= E\left(\frac{\hat{\lambda_1^*}}{\hat{\lambda_0^*}}\right) = \frac{n_0}{n_1}E\left(\frac{Y_1}{Y_0}\right) \\
&\approx \frac{n_0}{n_1}\left(\frac{\theta_1}{\theta_0} - 0 + \frac{\theta_0 * \theta_1}{\theta_0^3 + 3\theta_0^2 + \theta_0}\right) \\
&= \frac{n_0}{n_1}\left(\frac{\xi \exp(\omega)\overline{e_1}n_1}{\xi\overline{e_0}n_0} - 0 + \frac{\xi\overline{e_0}n_0 * \xi \exp(\omega)\overline{e_1}n_1}{(\xi\overline{e_0}n_0)^3 + 3(\xi\overline{e_0}n_0)^2 + (\xi\overline{e_0}n_0)}\right) \\
&= \exp(\omega)\frac{\overline{e_1}}{\overline{e_0}} + \frac{n_0 * \xi \exp(\omega)\overline{e_1}}{(\xi\overline{e_0}n_0)^2 + 3(\xi\overline{e_0}n_0) + 1} \\
&= \exp(\omega)\left[\frac{\overline{e_1}}{\overline{e_0}} + \frac{n_0 * \xi\overline{e_1}}{(\xi\overline{e_0}n_0)^2 + 3(\xi\overline{e_0}n_0) + 1}\right]. \qquad (4.5)
\end{aligned}$$

Therefore the bias, $E(\widehat{\exp(\omega^*)}) - \exp(\omega)$ is $\exp(\omega)\left[\left(\frac{\overline{e_1}}{\overline{e_0}} - 1\right) + \frac{n_0 * \xi\overline{e_1}}{(\xi\overline{e_0}n_0)^2 + 3(\xi\overline{e_0}n_0) + 1}\right]$.

A.2.3 Simulation Results: Partially Observed Follow-up Times

Table A.3: Bias and coverage of the rate ratio, $\exp(\mu)$, and between-study standard deviation, σ, using partially reported follow-up times: Simulation results for 20 primary studies as a function of relative follow-up in treatment arms. Percent bias [(estimated - true)/true $\times$ 100].

	RR=1				RR=0.5			
	$\sigma^2 = 0.01$		$\sigma^2 = 0.05$		$\sigma^2 = 0.01$		$\sigma^2 = 0.05$	
f	MCAR	MAR	MCAR	MAR	MCAR	MAR	MCAR	MAR
				RR: **Bias**				
0.9	0.21	0.21	3.41	3.39	5.44	5.42	3.52	3.52
0.95	2.35	2.37	-0.23	-0.22	-5.20	-5.16	-2.66	-2.64
1.0	2.50	2.50	-0.70	-0.71	0.76	0.76	-0.54	-0.54
1.05	0.33	0.31	2.59	2.59	-0.60	-0.56	4.84	4.82
1.1	3.68	3.69	4.47	4.44	1.18	1.02	1.45	1.50
1.2	1.25	1.28	7.82	7.85	2.52	2.55	8.30	8.26
				RR: **Coverage**				
0.9	0.995	0.997	0.992	0.993	0.935	0.923	0.997	0.997
0.95	0.971	0.969	1.000	1.000	0.918	0.918	0.992	0.990
1.0	0.951	0.954	1.000	1.000	0.997	0.997	0.999	0.999
1.05	0.991	0.995	0.992	0.993	0.997	0.998	0.998	0.997
1.1	0.904	0.906	0.990	0.987	0.997	0.996	1.000	1.000
1.2	0.983	0.981	0.891	0.899	0.967	0.963	0.881	0.886
				σ: **Bias**				
0.9	15.60	15.20	14.66	14.97	68.64	68.53	7.13	7.22
0.95	13.10	13.90	-8.45	-8.54	63.00	62.90	-24.68	-24.54
1.0	-4.60	-4.70	-7.28	-7.32	47.40	47.40	-10.08	-10.04
1.05	0.67	0.17	-34.69	-34.47	53.10	52.70	12.56	12.33
1.1	0.65	1.17	-12.19	-12.29	55.80	55.60	12.24	12.16
1.2	35.70	36.70	-3.18	-2.64	40.09	40.38	3.66	3.09
				σ: **Coverage**				
0.9	0.969	0.974	0.978	0.983	0.597	0.607	0.998	0.997
0.95	0.981	0.979	0.994	0.996	0.649	0.656	0.886	0.881
1.0	0.987	0.987	0.999	0.999	0.851	0.859	0.988	0.983
1.05	0.991	0.982	0.641	0.656	0.830	0.852	0.993	0.991
1.1	0.985	0.974	0.976	0.986	0.740	0.745	0.993	0.994
1.2	0.844	0.845	0.998	0.999	0.825	0.834	0.999	0.998

A.2.4 CRT Data Analysis: Ignoring Arm-Specific Follow-up for the 3 Studies Reporting Follow-Up

When ignoring the arm-specific follow-up information reported in the three studies, the posterior mean (95% credible interval) of the overall rate ratio was 0.69 (0.48, 0.93) using the most non-informative pair of priors, $\mu \sim$ Normal(0,1e06) and $\sigma \sim$ Half-Normal(0.26). This result also indicates a survival benefit of CRT-D compared to CRT-alone, but the benefit is slighly further from than null than when including the information (compare to 0.71 [0.49, 0.96] for using arm-specific follow-up). A similar comparison based on the most informative pair of priors, $\mu \sim$ Normal(0,2) and $\sigma \sim$ Half-Normal(0.03), the rate ratio is 0.69 [0.54, 0.86]. All other priors resulted in a similar overall rate ratio with 95% credible intervals that do not contain 1 (Table A.4). The posterior mean of the between-study standard deviation was estimated as 0.34 [0.03, 0.74], similar to that obtained using arm-specific follow-up (compare to 0.34 [0.08, 0.75])

Table A.4: CRT-D vs CRT-alone: posterior mean for the overall rate ratio and 95% credible intervals for 8 primary studies under a variety of prior distributions ignoring arm specific follow-up. ^{a}E(σ) = 0.14; ^{b}E(σ) = 0.35; ^{c}E(σ) = 0.41.

Prior for	Prior for overall underlying log rate ratio, μ		
between-study standard deviation, σ	Normal$(0, 2)$	Normal$(0, 10)$	Normal$(0, 1e06)$
Half-Normal$(0.03)^a$	0.69 (0.54, 0.86)	0.69 (0.54, 0.86)	0.69 (0.53, 0.86)
Uniform$(0, 0.7)^b$	0.69 (0.49, 0.92)	0.69 (0.49, 0.92)	0.68 (0.48, 0.93)
Half-Normal$(0.26)^c$	0.69 (0.49, 0.94)	0.69 (0.48, 0.96)	0.69 (0.48, 0.93)

A.3 A Maximum Likelihood Approach to Power Calculations for the Risk Difference in a Stepped Wedge Design for the Design of Left Ventricular Assist Devices for Destination Therapy

A.3.1 First and second derivatives

We want to take the derivative of this with respect to β_0, β_1 and τ^2. We will use the chain rule $\frac{\partial}{\partial t} f(g(t)) = f'(g(t))g'(t)$ where $f(t) = log(t)$ and $f'(t) = 1/t$ and $g() =$

$$\int (\frac{1}{\sqrt{2\pi}\tau} e^{-\frac{b^2}{2\tau^2}})^N \times (1-(\beta_0+b))^{Z_{00}} (\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}} db \tag{4.6}$$

We assume regularity conditions hold, that allow exchanging integration and differentiation in the calculation of $g'()$.

We first find the following integral, which we need for the chain rule above:

$$\frac{\partial}{\partial \beta_0} (\frac{1}{\sqrt{2\pi}\tau} e^{-\frac{b^2}{2\tau^2}})^N \times (1-(\beta_0+b))^{Z_{00}} (\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}}$$

Factor out constants:

$$(2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N \left(\frac{\partial}{\partial \beta_0} (1-(\beta_0+b))^{Z_{00}} (\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}}\right)$$

Product Rule with $(1-(\beta_0+b))^{Z_{00}}$ and $(\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}}$:

$$\begin{aligned}(2\pi)^{-N/2} (\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau})^N \Bigg[& (\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}} \frac{\partial}{\partial \beta_0} (1-(\beta_0+b))^{Z_{00}} \\ & + (1-(\beta_0+b))^{Z_{00}} \left(\frac{\partial}{\partial \beta_0} \left((\beta_0+b)^{Z_{01}} (1-(\beta_0+\beta_1+b))^{Z_{10}} (\beta_0+\beta_1+b)^{Z_{11}}\right)\right) \Bigg]\end{aligned}$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))^{Z_{00}}$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[Z_{00}(1-(\beta_0+b))^{Z_{00}-1}\left(\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))\right)(\beta_0+b)^{Z_{01}}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}}$$
$$\left(\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})\right)\Bigg]$$

Differentiate $\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))=-1$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(1-(\beta_0+b))^{Z_{00}}\left(\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})\right)\Bigg]$$

Product Rule with $(1-(\beta_0+\beta_1+b))^{Z_{10}}$ and $(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(1-(\beta_0+b))^{Z_{00}}\Bigg((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}\frac{\partial}{\partial\beta_0}((1-(\beta_0+\beta_1+b))^{Z_{10}})$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}})\Bigg)\Bigg]$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(1-(\beta_0+\beta_1+b))^{Z_{10}}$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(1-(\beta_0+b))^{Z_{00}}\Bigg((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$\left(\frac{\partial}{\partial\beta_0}((1-(\beta_0+\beta_1+b)))\right)+(1-(\beta_0+\beta_1+b))^{Z_{10}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}})\Bigg)\Bigg]$$

Differentiate $\frac{\partial}{\partial\beta_0}(1-(\beta_0+\beta_1+b))=-1$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(1-(\beta_0+b))^{Z_{00}}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}})\Bigg)\Bigg]$$

Product Rule with $(\beta_0+b)^{Z_{01}}$ and $(\beta_0+\beta_1+b)^{Z_{11}}$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(1-(\beta_0+b))^{Z_{00}}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\Bigg((\beta_0+\beta_1+b)^{Z_{11}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}})$$
$$+(\beta_0+b)^{Z_{01}}\frac{\partial}{\partial\beta_0}((\beta_0+\beta_1+b)^{Z_{11}}))\Bigg)\Bigg]$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(\beta_0+b))^{Z_{01}}$ and differentiate $\frac{\partial}{\partial\beta_0}(\beta_0+b)=1$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(1-(\beta_0+b))^{Z_{00}}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\left((\beta_0+\beta_1+b)^{Z_{11}}Z_{01}(\beta_0+b)^{Z_{01}-1}\right.$$
$$\left.+(\beta_0+b)^{Z_{01}}\frac{\partial}{\partial\beta_0}((\beta_0+\beta_1+b)^{Z_{11}}))\right)\Bigg]$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(\beta_0+\beta_1+b))^{Z_{11}}$ and differentiate $\frac{\partial}{\partial\beta_0}(\beta_0+\beta_1+b)=1$:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(1-(\beta_0+b))^{Z_{00}}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\left((\beta_0+\beta_1+b)^{Z_{11}}Z_{01}(\beta_0+b)^{Z_{01}-1}\right.$$
$$\left.+(\beta_0+b)^{Z_{01}}Z_{11}(\beta_0+\beta_1+b)^{Z_{11}-1}\right)\Bigg)\Bigg]$$

Simplifying the result is:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{01}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Bigg]$$

Now we use this result, adding the integral over b back in, combined with the chain rule to obtain the final result:

$$\frac{1}{\int(\frac{1}{\sqrt{2\pi}\tau}e^{-\frac{b^2}{2\tau^2}})^N\times(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}db}\times$$
$$\int(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{01}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Bigg]db \tag{4.7}$$

We first find the following integral, which we need for the chain rule above:

$$\frac{\partial}{\partial\beta_1}(\frac{1}{\sqrt{2\pi}\tau}e^{-\frac{b^2}{2\tau^2}})^N\times(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$

The derivation is similar to the previous derivative with respect to β_0, but we have some simplification because only 2 of the terms have β_1 and the result is:

$$(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N (1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}$$
$$\left[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}+Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\right]$$

Now we use this result, adding the integral over b back in, combined with the chain rule to obtain the final result:

$$\frac{1}{\int(\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{b^2}{2\tau^2}})^N\times(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}db}\times$$
$$\int(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N (1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}\Big[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]db \tag{4.8}$$

We first find the following integral, which we need for the chain rule above:

$$\frac{\partial}{\partial\tau^2}(\frac{1}{\sqrt{2\pi}\tau}e^{-\frac{b^2}{2\tau^2}})^N\times(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$

Chain Rule for $(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^N$ and factor out constants with respect to τ^2:

$$(\frac{1}{2\pi})^{N/2}((1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}$$
$$\frac{\partial}{\partial\tau^2}(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})$$

Product Rule with $e^{-\frac{b^2}{2\tau^2}}$ and $\frac{1}{\sqrt{\tau^2}}$:

$$(\frac{1}{2\pi})^{N/2}((1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}$$
$$\left(\frac{1}{\sqrt{\tau^2}}\frac{\partial}{\partial\tau^2}(e^{-\frac{b^2}{2\tau^2}})+e^{-\frac{b^2}{2\tau^2}}\frac{\partial}{\partial\tau^2}(\frac{1}{\sqrt{\tau^2}})\right)$$

Chain Rule for $e^{-\frac{b^2}{2\tau^2}}$ and differentiate $\frac{\partial}{\partial\tau^2}(-\frac{b^2}{2\tau^2}) = \frac{b^2}{2(\tau^2)^2}$ and differentiate $\frac{\partial}{\partial\tau^2}(\frac{1}{\sqrt{\tau^2}}) = -\frac{1}{2(\tau^2)^{3/2}}$:

$$(\frac{1}{2\pi})^{N/2}((1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}$$
$$\left(\frac{b^2}{2(\tau^2)^{5/2}}(e^{-\frac{b^2}{2\tau^2}})+e^{-\frac{b^2}{2\tau^2}}(-\frac{1}{2(\tau^2)^{3/2}})\right)$$

Simplifying the result is:

$$(\frac{1}{2\pi})^{N/2}((1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}$$
$$\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)$$

Now we use this result, adding the integral over b back in, combined with the chain rule to obtain the final result:

$$\frac{1}{\int(\frac{1}{\sqrt{2\pi\tau}}e^{-\frac{b^2}{2\tau^2}})^N\times(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}db}\times$$
$$\int(\frac{1}{2\pi})^{N/2}((1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}$$
$$\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)db \tag{4.9}$$

We need the second derivatives for the expected Fisher information matrix, if we prefer to use this to estimate the asymptotic variance of $\hat{\beta}_1$ instead of the outer product of gradients, which only requires the first derivatives of the log likelihood.

The results from the first derivatives are all of the form $f(t)/g(t)$. We will use the quotient rule: $\frac{\partial}{\partial t}\frac{f(t)}{g(t)} = \frac{f'(t)g(t)-f(t)g'(t)}{[g(t)]^2}$. Both $f(t)$ and $g(t)$ are integrals, so we again assume regularity conditions hold, that allow exchanging integration and differentiation in order to calculate $f'(t)$ and $g'(t)$.

Note that $g(t) = \int (\frac{1}{\sqrt{2\pi}\tau} e^{-\frac{b^2}{2\tau^2}})^N \times (1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}} db$ (Eq (4.6)) for each case. This is the integral we just worked with for the first derivatives. Hence, we have the results for $g'(t)$ in the previous section and can plug these in. Hence, we will focus on taking derivatives inside the integral for the numerators.

Derivatives inside the integral of $(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N (-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$**, the first term of the numerator of** $\frac{\partial}{\partial\beta_0} l(\boldsymbol{\beta}, \tau^2)$**:**

with respect to β_0

$$\frac{\partial}{\partial\beta_0}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N (-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Factor out constants:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N \left(\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}\right.$$
$$\left.(\beta_0+\beta_1+b)^{Z_{11}}\right)$$

Product Rule with $(1-(\beta_0+b))^{Z_{00}-1}$ and $(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N \left[(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))^{Z_{00}-1}\right.$$
$$\left.+(1-(\beta_0+b))^{Z_{00}-1}\left(\frac{\partial}{\partial\beta_0}\left((\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}\right)\right)\right]$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))^{Z_{00}-1}$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}\left(\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))\right)(\beta_0+b)^{Z_{01}}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}$$
$$\left(\frac{\partial}{\partial\beta_0}\left((\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}\right)\right)\Bigg]$$

Differentiate $\frac{\partial}{\partial\beta_0}(1-(\beta_0+b))=-1$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}\left(\frac{\partial}{\partial\beta_0}\left((\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}\right)\right)\Bigg]$$

Product Rule with $(1-(\beta_0+\beta_1+b))^{Z_{10}}$ and $(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}\Bigg((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}\frac{\partial}{\partial\beta_0}((1-(\beta_0+\beta_1+b))^{Z_{10}})+$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}})\Bigg)\Bigg]$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(1-(\beta_0+\beta_1+b))^{Z_{10}}$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}\Bigg((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\frac{\partial}{\partial\beta_0}((1-(\beta_0+\beta_1+b))))+(1-(\beta_0+\beta_1+b))^{Z_{10}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}})\Bigg)\Bigg]$$

Differentiate $\frac{\partial}{\partial\beta_0}(1-(\beta_0+\beta_1+b))=-1$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}})\Bigg)\Bigg]$$

Product Rule with $(\beta_0+b)^{Z_{01}}$ and $(\beta_0+\beta_1+b)^{Z_{11}}$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\left((\beta_0+\beta_1+b)^{Z_{11}}\frac{\partial}{\partial\beta_0}((\beta_0+b)^{Z_{01}})+(\beta_0+b)^{Z_{01}}\frac{\partial}{\partial\beta_0}((\beta_0+\beta_1+b)^{Z_{11}})\right)\Bigg)\Bigg]$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(\beta_0+b))^{Z_{01}}$ and differentiate $\frac{\partial}{\partial\beta_0}(\beta_0+b)=1$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\left((\beta_0+\beta_1+b)^{Z_{11}}Z_{01}(\beta_0+b)^{Z_{01}-1}+(\beta_0+b)^{Z_{01}}\frac{\partial}{\partial\beta_0}((\beta_0+\beta_1+b)^{Z_{11}})\right)\Bigg)\Bigg]$$

Chain Rule for $\frac{\partial}{\partial\beta_0}(\beta_0+\beta_1+b))^{Z_{11}}$ and differentiate $\frac{\partial}{\partial\beta_0}(\beta_0+\beta_1+b)=1$:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Bigg[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(1-(\beta_0+b))^{Z_{00}-1}\Bigg(-(\beta_0+b)^{Z_{01}}(\beta_0+\beta_1+b)^{Z_{11}}Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$+(1-(\beta_0+\beta_1+b))^{Z_{10}}\left((\beta_0+\beta_1+b)^{Z_{11}}Z_{01}(\beta_0+b)^{Z_{01}-1}+(\beta_0+b)^{Z_{01}}Z_{11}(\beta_0+\beta_1+b)^{Z_{11}-1}\right)\Bigg)\Bigg]$$

Simplifying the result is:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Big[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{01}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]$$

Hence, derivative with respect to β_0 of the first term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int -Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Big[-(Z_{00}-1)(1-(\beta_0+b))^{Z_{00}-2}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{01}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]db \qquad (4.10)$$

with respect to β_1

$$\frac{\partial}{\partial\beta_1}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

The derivation is similar to the previous derivative with respect to β_0, but we have some simplification because only 2 of the terms have β_1 and the result is:

$$-Z_{00}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}\Big[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]$$

Hence, derivative with respect to β_1 of the first term of the numerator of $\frac{\partial}{\partial \beta_0} l(\boldsymbol{\beta}, \tau^2)$ is

$$\int -Z_{00}(2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}} \Big[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}} + Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1} \Big] db \tag{4.11}$$

with respect to τ^2

$$\frac{\partial}{\partial \tau^2}((2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Chain Rule for $(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^N$ and factor out constants with respect to τ^2:

$$(\frac{1}{2\pi})^{N/2}((-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\frac{\partial}{\partial \tau^2}(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})$$

Product Rule with $e^{-\frac{b^2}{2\tau^2}}$ and $\frac{1}{\sqrt{\tau^2}}$:

$$(\frac{1}{2\pi})^{N/2}((-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{1}{\sqrt{\tau^2}}\frac{\partial}{\partial \tau^2}(e^{-\frac{b^2}{2\tau^2}}) + e^{-\frac{b^2}{2\tau^2}}\frac{\partial}{\partial \tau^2}(\frac{1}{\sqrt{\tau^2}})\right)$$

Chain Rule for $e^{-\frac{b^2}{2\tau^2}}$ and differentiate $\frac{\partial}{\partial \tau^2}(-\frac{b^2}{2\tau^2}) = \frac{b^2}{2(\tau^2)^2}$ and differentiate $\frac{\partial}{\partial \tau^2}(\frac{1}{\sqrt{\tau^2}}) = -\frac{1}{2(\tau^2)^{3/2}}$:

$$(\frac{1}{2\pi})^{N/2}((-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2}{2(\tau^2)^{5/2}}(e^{-\frac{b^2}{2\tau^2}}) + e^{-\frac{b^2}{2\tau^2}}(-\frac{1}{2(\tau^2)^{3/2}})\right)$$

Simplifying the result is:

$$(\frac{1}{2\pi})^{N/2}((-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)$$

Hence, derivative with respect to τ^2 of the first term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int(\frac{1}{2\pi})^{N/2}((-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)db \tag{4.12}$$

Derivatives inside the integral of $(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$**, the second term of the numerator of** $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$**:**

with respect to β_0

$$\frac{\partial}{\partial\beta_0}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Similar to the derivative with respect to β_0 of the first term of the numerator, the result is:

$$Z_{01}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N$$
$$\Big[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(Z_{01}-1)(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-2}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]$$

Hence, derivative with respect to β_0 of the second term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int Z_{01}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N$$
$$\Big[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+(Z_{01}-1)(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-2}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]db \qquad (4.13)$$

with respect to β_1

$$\frac{\partial}{\partial\beta_1}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Similar to the derivative with respect to β_1 of the first term of the numerator, the result is:

$$Z_{01}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}\Big[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]$$

Hence, derivative with respect to β_1 of the second term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int Z_{01}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}\Big[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]db \qquad (4.14)$$

with respect to τ^2

$$\frac{\partial}{\partial\tau^2}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Similar to the derivative with respect to τ^2 of the first term of the numerator, the result is:

$$(\frac{1}{2\pi})^{N/2}((Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)$$

Hence, derivative with respect to τ^2 of the second term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int(\frac{1}{2\pi})^{N/2}((Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)db \tag{4.15}$$

Derivatives inside the integral of $(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(-Z_{10})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$**, the third term of the numerator of** $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$**: with respect to** β_0

$$\frac{\partial}{\partial\beta_0}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(-Z_{10})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Similar to the derivative with respect to β_0 of the first term of the numerator, the result is:

$$-Z_{10}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Big[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-(Z_{10}-1)(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-2}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]$$

Hence, derivative with respect to β_0 of the third term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int -Z_{10}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Big[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+(Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$-(Z_{10}-1)(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-2}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+Z_{11}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]db \qquad (4.16)$$

with respect to β_1

$$\frac{\partial}{\partial\beta_1}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(-Z_{10})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Similar to the derivative with respect to β_1 of the first term of the numerator, the result is:

$$-Z_{10}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}\Big[-(Z_{10}-1)(1-(\beta_0+\beta_1+b))^{Z_{10}-2}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]$$

Hence, derivative with respect to β_1 of the third term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int -Z_{10}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}\Big[-(Z_{10}-1)(1-(\beta_0+\beta_1+b))^{Z_{10}-2}$$
$$(\beta_0+\beta_1+b)^{Z_{11}}+Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}-1}\Big]db \qquad (4.17)$$

with respect to τ^2

$$\frac{\partial}{\partial\beta_1}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(-Z_{10})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}})$$

Similar to the derivative with respect to τ^2 of the first term of the numerator, the result is:

$$(\frac{1}{2\pi})^{N/2}((-Z_{10})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)$$

Hence, derivative with respect to τ^2 of the third term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int(\frac{1}{2\pi})^{N/2}((-Z_{10})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)db \tag{4.18}$$

Derivatives inside the integral of $(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{11})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}$**, the fourth term of the numerator of** $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$**:**

with respect to β_0

$$\frac{\partial}{\partial\beta_0}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{11})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}-1}$$

Similar to the derivative with respect to β_0 of the first term of the numerator, the result is:

$$Z_{11}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Big[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}-1}+Z_{01}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}-1}$$
$$+(Z_{11}-1)(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-2}\Big]$$

Hence, derivative with respect to β_0 of the fourth term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int Z_{11}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N\Big[-Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}-1}+Z_{01}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}$$
$$-Z_{10}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}-1}$$
$$+(Z_{11}-1)(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-2}\Big]db \quad (4.19)$$

with respect to β_1

$$\frac{\partial}{\partial\beta_1}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{11})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}$$

Similar to the derivative with respect to β_1 of the first term of the numerator, the result is:

$$Z_{11}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}\Big[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}-1}+(Z_{11}-1)(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-2}\Big]$$

Hence, derivative with respect to β_1 of the fourth term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int Z_{11}(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}\Big[-Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}-1}+(Z_{11}-1)(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-2}\Big]db \quad (4.20)$$

with respect to τ^2

$$\frac{\partial}{\partial\tau^2}((2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(Z_{11})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}$$

Similar to the derivative with respect to τ^2 of the first term of the numerator, the result is:

$$(\frac{1}{2\pi})^{N/2}((Z_{11})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)$$

Hence, derivative with respect to τ^2 of the fourth term of the numerator of $\frac{\partial}{\partial\beta_0}l(\boldsymbol{\beta},\tau^2)$ is

$$\int(\frac{1}{2\pi})^{N/2}((Z_{11})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1})$$
$$N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)db \tag{4.21}$$

Derivatives inside the integral of $(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(-Z_{10})(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$**, the first term of the numerator of** $\frac{\partial}{\partial\beta_1}l(\boldsymbol{\beta},\tau^2)$**:**

These have been previously derived.

Derivatives inside the integral of $(2\pi)^{-N/2}\left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau}\right)^N(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(Z_{11})(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1}$**, the second term of the numerator of** $\frac{\partial}{\partial\beta_1}l(\boldsymbol{\beta},\tau^2)$**:**

These have been previously derived.

Derivatives inside the integral of
$(\frac{1}{2\pi})^{N/2}((1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}})N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1}\left(\frac{b^2e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}}-\frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}}\right)$**,**
the numerator of $\frac{\partial}{\partial\tau^2}l(\boldsymbol{\beta},\tau^2)$**:**

with respect to β_0

The result including the integral is:

$$\int (\frac{1}{2\pi})^{N/2} N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1} \left(\frac{b^2 e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}} - \frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}} \right) \Big[- Z_{00}(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}} + Z_{01}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}} - Z_{10}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}} + Z_{11}(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1} \Big] db \tag{4.22}$$

with respect to β_1

The result including the integral is:

$$\int (\frac{1}{2\pi})^{N/2} N(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1} \left(\frac{b^2 e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}} - \frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}} \right) (1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}$$
$$\Big[- Z_{10}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}}$$
$$+ Z_{11}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1} \Big] db \tag{4.23}$$

with respect to τ^2

The result including the integral is:

$$\int (1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}}(\frac{1}{2\pi})^{N/2}$$
$$N \Big[\frac{1}{4\tau^{2^4}} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}} \right)^N \Big(b^4 N - b^2\tau^2(2N+4) + \tau^{2^2}(N+2) \Big) \Big] db \tag{4.24}$$

Combining the results

To get the second derivatives we noted that we need to use the quotient rule: $\frac{\partial}{\partial t}\frac{f(t)}{g(t)} = \frac{f'(t)g(t)-f(t)g'(t)}{[g(t)]^2}$. Below, I note which equations derived above need to be plugged into this rule to get the second derivatives.

For $\frac{\partial}{\partial \beta_0^2}$:

$$f = \Bigg[\int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (-Z_{10})(1-(\beta_0+b))^{Z_{00}}$$
$$(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (Z_{11})$$
$$(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1} db \Bigg]$$

$f' =$ (4.10) + (4.13) + (4.16) + (4.19)

$g =$ (4.6)

$g' =$ (4.7)

For $\frac{\partial}{\partial \beta_0 \beta_1}$:

$$f = \Bigg[\int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (-Z_{10})(1-(\beta_0+b))^{Z_{00}}$$
$$(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (Z_{11})$$
$$(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1} db \Bigg]$$

$f' = (4.11) + (4.14) + (4.17) + (4.20)$

$g = (4.6)$

$g' = (4.8)$

For $\frac{\partial}{\partial \beta_0 \tau^2}$:

$$f = \Bigg[\int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (-Z_{00})(1-(\beta_0+b))^{Z_{00}-1}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}$$
$$(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (Z_{01})(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}-1}$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (-Z_{10})(1-(\beta_0+b))^{Z_{00}}$$
$$(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}-1}(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (Z_{11})$$
$$(1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1} db \Bigg]$$

$f' = (4.12) + (4.15) + (4.18) + (4.21)$

$g = (4.6)$

$g' = (4.9)$

For $\frac{\partial}{\partial \beta_1^2}$:

$$f = \Bigg[\int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(-Z_{10})(1-(\beta_0+\beta_1+b))^{Z_{10}-1}$$
$$(\beta_0+\beta_1+b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (1-(\beta_0+b))^{Z_{00}}(\beta_0+b)^{Z_{01}}(Z_{11})$$
$$(1-(\beta_0+\beta_1+b))^{Z_{10}}(\beta_0+\beta_1+b)^{Z_{11}-1} db \Bigg]$$

$f' =$ (4.17) + (4.20)

$g =$ (4.6)

$g' =$ (4.8)

For $\frac{\partial}{\partial \beta_1 \tau^2}$:

$$f = \Bigg[\int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (1 - (\beta_0 + b))^{Z_{00}} (\beta_0 + b)^{Z_{01}} (-Z_{10}) (1 - (\beta_0 + \beta_1 + b))^{Z_{10}-1}$$
$$(\beta_0 + \beta_1 + b)^{Z_{11}} db + \int (2\pi)^{-N/2} \left(\frac{e^{-\frac{b^2}{2\tau^2}}}{\tau} \right)^N (1 - (\beta_0 + b))^{Z_{00}} (\beta_0 + b)^{Z_{01}} (Z_{11})$$
$$(1 - (\beta_0 + \beta_1 + b))^{Z_{10}} (\beta_0 + \beta_1 + b)^{Z_{11}-1} db \Bigg]$$

$f' =$ (4.18) + (4.21)

$g =$ (4.6)

$g' =$ (4.9)

For $\frac{\partial}{\partial \tau^2 2}$:

$$f = \int (\frac{1}{2\pi})^{N/2} ((1 - (\beta_0 + b))^{Z_{00}} (\beta_0 + b)^{Z_{01}} (1 - (\beta_0 + \beta_1 + b))^{Z_{10}} (\beta_0 + \beta_1 + b)^{Z_{11}})$$
$$N (\frac{e^{-\frac{b^2}{2\tau^2}}}{\sqrt{\tau^2}})^{N-1} \left(\frac{b^2 e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{5/2}} - \frac{e^{-\frac{b^2}{2\tau^2}}}{2(\tau^2)^{3/2}} \right) db$$

$f' =$ (4.24)

$g =$ (4.6)

$g' =$ (4.9)

A.3.2 Computational Details

The derivatives are quite complex and involve integrating over the distribution of random effects. For reasonable design parameters for the SWD regarding the number of clusters I, steps J, number of people sampled N, ICC and baseline proportion and risk difference, we encountered difficulties computing these integrals as numeric overflow occurs because the values of the integrands are extremely large near b=0 for the random effect. The standard precision utilized by R is double precision and the largest value allowed before it is labeled as called infinity is 1.797×10^{308}. For many of our examples within the SWD parameter space, we were beyond this capacity. For a SWD with $J = 5$ and 100 people sampled at each step, $N = 500$ and for rare outcomes, the value of the integrand was larger than that allowed in double precision.

We were able to work in quadruple precision rather than double precision, utilizing the R package Rmpfr, where MPFR is acronym for "Multiple Precision Floating-Point Reliably". In Rmpfr, we are allowed to increase the precision from double precision by increasing the number of bits. If we set the number of bits to be 53, we would use double precision. This R package calls to GNU MPFR, a portable C library for arbitrary-precision binary floating-point computation with correct rounding, based on GNU Multi-Precision Library.

The Rmpfr package uses the Romberg algorithm for integration. Note that in scientific notation all numbers are written in the form $M \times 10^E$ or MeE. The exponent is E and M is called the mantissa. When dealing with these extremely large integrals, set the convergence criteria as follows: (1) when the exponent from the current order E is equal to the exponent from the previous order E_{-1} *and* (2) when the absolute value of the difference between the mantissa for the current and previous order is less than 10^{-8} or $|M - M_{-1}| < 10^{-8}$. When this occurs we stop at the order that satisfies these criteria and report the value of the integral.

In addition, we considered Monte Carlo integration to evaluate the integrals that make

up the derivatives. The results of Monte Carlo integration were very similar to those obtained by Romberg integration even for relatively large N, but computing time was significantly decreased for the Romberg method (results not shown).

References

(2013). "The pcori (patient-centered outcomes research institute) methodology report," URL http://www.pcori.org/research-we-support/research-methodology-standards.

(2014). "What is comparative effectiveness research," URL http://effectivehealthcare.ahrq.gov/ index.cfm/what-is-comparative-effectiveness-research1/.

Abadie, A. and G. W. Imbens (2006). "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74, 235–267.

Adlbrecht, C., M. Hulsmann, M. Gwechenberger, G. Strunk, C. Khazen, F. Wiesbauer, M. Elhenicky, S. Neuhold, T. Binder, G. Maurer, I. M. Lang, and R. Pacher (2009). "Outcome after device implantation in chronic heart failure is dependent on concomitant medical treatment," *European Journal of Clinical Investigation*, 39, 1073–1081.

Ahern, J., A. Hubbard, and S. Galea (2009). "Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods," *American Journal of Epidemiology*, 169, 1140–1147.

Auricchio, A., M. Metra, M. Gasparini, B. Lamp, C. Klersy, A. Curnis, C. Fantoni, E. Gronda, and J. Vogt (2007). "Long-term survival of patients with heart failure and ventricular conduction delay treated with cardiac resynchronization therapy," *The American Journal of Cardiology*, 99, 232–238.

Austin, P. C. and M. M. Mamdani (2006). "A comparison of propensity score methods: A casestudy estimating the effectiveness of postami statin use," *Statistics in Medicine*, 25, 2084–2106.

Bai, R., L. D. Biase, C. Elayi, C. K. Ching, C. Barrett, K. Philipps, P. Lim, D. Patel, T. Callahan, D. O. Martin, M. Arruda, R. A. Schweikert, W. I. Saliba, B. Wilkoff, and A. Natale (2008). "Mortality of heart failure patients after cardiac resynchronization therapy: Identification of predictors," *Journal of Cardiovascular Electrophysiology*, 19, 1259–1265.

Bristow, M. R., L. A. Saxon, J. Boehmer, S. Krueger, D. A. Kass, T. D. Marco, P. Carson, L. DiCarlo, D. DeMets, B. G. White, D. W. DeVries, and A. M. Feldman (2004). "Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure," *The New England Journal of Medicine*, 350, 2140–2150.

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Sturmer (2006). "Variable selection for propensity score models," *American Journal of Epidemiology*, 163, 1149–1156.

Brown, C. A. and R. J. Lilford (2006). "The stepped wedge trial design: a systematic review," *BMC Medical Research Methodology*, 6.

Bushman, B. J. and M. C. Wang (1996). "A procedure for combining sample standardized mean differences and vote counts to estimate the population standardized mean difference in fixed effects models," *Psychological Methods*, 1, 66–80.

Carlin, B. P. and T. A. Louis (2001). *Bayes and empirical Bayes methods for data analysis*, London: Chapman and Hall, 2 edition.

Cousens, S., J. Hargreaves, C. Bonell, B. Armstrong, J. Thomas, and B. R. Kirkwood (2011). "Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference," *J Epidemiol Community Health*, 65, 576–581.

Donner, A. and N. Klar (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*, London: Arnold.

Droitcour, J., G. Silberman, and E. Chelimsky (1993). "Cross-design synthesis: A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases," *International Journal of Technology Assessment in Health Care*, 9, 440–449.

Dunlay, S. M., S. J. Park, L. D. Joyce, R. C. Daly, J. M. Stulak, S. M. McNallan, V. L. Roger, and S. S. Kushwaha (2014). "Frailty and outcomes after implantation of left ventricular assist device as destination therapy," *J. Heart Lung Transplant*, 33, 359–365.

Ermis, C., K. G. Lurie, A. X. Zhu, J. Collins, L. Vanheel, S. Sakaguchi, F. Lu, S. Pham, and D. G. Benditt (2004). "Biventricular implantable cardioverter defibrillators improve survival compared with biventricular pacing alone in patients with severe left ventricular dysfunction," *Journal of Cardiovascular Electrophysiology*, 15, 862–866.

Feldman, A. M., G. de Lissovoy, M. R. Bristow, L. A. Saxon, T. D. Marco, D. A. Kass, J. Boehmer, S. Singh, D. J. Whellan, P. Carson, A. Boscoe, T. M. Baker, and M. R. Gunderman (2005). "Cost effectiveness of cardiac resynchronization therapy in the comparison of medical therapy, pacing, and defibrillation in heart failure (companion) trial," *Journal of the American College of Cardiology*, 46, 2311–2321.

Greenland, S. and J. M. Robins (1986). "Identifiability, exchangeability, and epidemiological confounding," *International Journal of Epidemiology*, 15, 413–419.

Gu, X. and P. R. Rosenbaum (1993). "Comparison of multivariate matching methods: Structures, distances, and algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420.

Haviland, A., D. Nagin, and P. R. Rosenbaum (2007). "Combining propensity score matching and group-based trajectory analysis in an observational study," *Psychological Methods*, 12, 247–267.

Heckman, J. J., H. Hidehiko, and P. Todd (1997). "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme," *Review of Economic Studies*, 64, 605–654.

Hintze, J. (2008). *PASS 2008*, NCSS, LLC, Kaysville, Utah, USA, pass 2008 edition.

Holland, P. W. (1986). "Statistics and causal inference," *Journal of the American Statistical Association*, 81, 945–960.

Horvitz, D. G. and D. J. Thompson (1952). "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47, 663–685.

Hussey, M. A. and J. P. Hughes (2007). "Design and analysis of stepped wedge cluster randomized trials," *Contemporary Clinical Trials*, 28, 182–191.

Imbens, G. W. (2000). "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

Imbens, G. W. and J. D. Angrist (1994). "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Kang, J. D. Y. and J. L. Schafer (2007). "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 22, 523–539.

Kirklin, J. K., D. C. Naftel, R. L. Kormos, L. W. Stevenson, F. D. Pagani, M. A. Miller, J. T. Baldwin, and J. B. Young (2012). "The fourth intermacs annual report: 4,000 implants and counting," *J. Heart Lung Transplant*, 31, 117–126.

Kirklin, J. K., D. C. Naftel, R. L. Kormos, L. W. Stevenson, F. D. Pagani, M. A. Miller, K. L. Ulisney, J. T. Baldwin, and J. B. Young (2011). "Third intermacs annual report: the evolution of destination therapy in the united states," *J. Heart Lung Transplant.*, 30, 115–123.

Konstam, M. A., I. Pina, J. Lindenfeld, and M. Packer (2003). "A device is not a drug," *Journal of Cardiac Failure*, 9, 155–157.

Lee, B., J. Lessler, and E. A. Stuart (2009). "Improving propensity score weighting using machine learning," *Statistics in Medicine*, 29, 337–346.

Lu, B., E. Zanutto, R. Hornik, and P. R. Rosenbaum (2001). "Matching with doses in an observational study of a media campaign against drug abuse," *Journal of the American Statistical Association*, 96, 1245–1253.

Lumley, T. (2002). "Network meta-analysis for indirect treatment comparisons," *Statistics in Medicine*, 21, 2313–2324.

Lunceford, J. K. and M. Davidian (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, 23, 2937–2960.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). "Winbugs - a bayesian modelling framework: concepts, structure, and extensibility." *Statistics and Computing*, 10, 325–337.

Mauri, L., T. S. Silbaugh, P. Garg, R. E. Wolf, K. Zelevinsky, A. Lovett, M. R. Varma, Z. Zhou, and S.-L. T. Normand (2008). "Drug-eluting or bare-metal stents for acute myocardial infarction," *New England Journal of Medicine*, 359, 1330–1342.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). "Propensity score estimation with boosted regression for evaluating causal effects in observational studies," *Psychological Methods*, 9, 403–425.

Mdege, N. D., M. S. Man, C. A. Taylor Nee Brown, and D. J. Torgerson (2011). "Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation," *J Clin Epidemiol*, 64, 936–949.

Neaton, J. D., S. L. Normand, A. Gelijns, R. C. Starling, D. L. Mann, and M. A. Konstam (2007). "Designs for mechanical circulatory support device studies," *J. Card. Fail.*, 13, 63–74.

Newey, W. K. and D. McFadden (1994). *Handbook of econometrics*, volume IV, Elsevier Science.

Normand, S.-L. T., M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil (2001). "Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores," *Journal of clinical epidemiology*, 54, 387–398.

Pappone, C., G. Vicedomini, G. Augello, P. Mazzone, S. Nardi, and S. Rosanio (2003). "Combining electrical therapies for advanced heart failure: The milan experience with biventricular pacing–defibrillation backup combination for primary prevention of sudden cardiac death," *The American Journal of Cardiology*, 91, 74F–80F.

Parmar, M. K. B., V. Torri, and L. Stewart (1998). "Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints," *Statistics in Medicine*, 17, 2815–2834.

Potter, F. J. (1993). "The effect of weight trimming on nonlinear survey estimates," in *Proceedings of the Section on Survey Research Methods of American Statistical Association*, San Francisco, CA: American Statistical Association.

Rassen, J. A., A. A. Shelat, J. Myers, R. J. Glynn, K. J. Rothman, and S. Schneeweiss (2012). "One-to-many propensity score matching in cohort studies," *Pharmacoepidemiology and Drug Safety*, 21, 69–80.

Robins, J. M. (1986). "A new approach to causal inference in mortality studies with sustained exposure periods: Application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

Robins, J. M., M. Hernan, and B. Brumback (2000). "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11, 550–560.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846–866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). "Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, 90, 106–121.

Rose, E. A., A. C. Gelijns, A. J. Moskowitz, D. F. Heitjan, L. W. Stevenson, W. Dembitsky, J. W. Long, D. D. Ascheim, A. R. Tierney, R. G. Levitan, J. T. Watson, P. Meier, N. S. Ronan, P. A. Shapiro, R. M. Lazar, L. W. Miller, L. Gupta, O. H. Frazier, P. Desvigne-Nickens, M. C. Oz, and V. L. Poirier (2001). "Long-term use of a left ventricular assist device for end-stage heart failure," *New England Journal of Medicine*, 345, 1435–1443.

Rose, S. (2013). "Mortality risk score prediction in an elderly population using machine learning," *Am J Epidemiol*, 177, 443–452.

Rosenbaum, P. R. (1987). "The role of a second control group in an observational study," *Statistical Science*, 2, 292–316.

Rosenbaum, P. R. (2002). *Observational Studies*, New York: Springer, 2 edition.

Rosenbaum, P. R. and D. B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and D. B. Rubin (1984). "Reducing bias in observational studies using subclassification on the propensity score," *Journal of the American Statistical Association*, 79, 516–524.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of educational Psychology*, 66, 688–701.

Rubin, D. B. (2007). "The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials," *Statistics in Medicine*, 26, 20–36.

Schuchert, A., C. Muto, T. Maounis, R. Frank, E. Boulogne, A. Polauck, and L. Padeletti (2013). "Lead complications, device infections, and clinical outcomes in the first year after implantation of cardiac resynchronization therapy-defibrillator and cardiac resynchronization therapy-pacemaker," *Europace*, 15, 71–76.

Sekhon, J. S. (2008). "The neyman-rubin model of causal inference and estimation via matching methods," *The Oxford Handbook of Political Methodology*, 271–299.

Setoguchi, S., S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook (2008). "Evaluating uses of data mining techniques in propensity score estimation: A simulation study," *Pharmacoepidemiology and Drug Safety*, 17, 546–555.

Slaughter, M. S., J. G. Rogers, C. A. Milano, S. D. Russell, J. V. Conte, D. Feldman, B. Sun, A. J. Tatooles, R. M. Delgado, J. W. Long, T. C. Wozniak, W. Ghumman, D. J. Farrar, O. H. Frazier, M. Sobieski, C. Gallagher, P. Pappas, M. Silver, A. Lodge, L. Blue, A. Shah, D. Yuh, S. Ullrich, D. Dordunoo, D. Rivard, B. Kar, B. Radovancevic, I. Gregoric, A. Civitello, E. Massin, C. Gemmato, M. Jafar, R. Bogaev, F. Smart, J. Sirak, S. Sudhaker, T. Yanssens, B. Reid, S. Horton, D. Renland, J. Revenaugh, M. Eidson, M. Turrentine, S. Becka, D. Dean, S. Murali, G. Magovern, S. Bailey, G. Sokos, L. Kernickey, N. Moazami, G. Ewald, K. Shelton, D. Anderson, I. Wang, E. Garrett, T. Edwards, R. Carter, C. Porter, P. Shekar, G. Couper, M. Givertz, S. Kelly, E. Raines, K. Miller, L. McClement-Green, E. Haeusslein, G. J. Avery, P. Brandenhoff, J. Carnam, T. Oka, R. Courville, N. Smedira, R. Starling, J. Navia, G. Gonzalez, T. Mihaljevic, L. Teague, Y. Naka, K. Idrissi, A. Stewart, D. Vega, A. Smith, R. Laskar, J. Thompson, J. Entwistle, H. Eisen, S. Hankins, T. Metzger, R. Brewer, B. Czerska, C. Williams, B. Braxton, W. Pae, J. Boehmer, T. Stephensen, M. Lazar, A. Myers, M. Acker, M. Jessup, R. Morris, S. Desai, M. O'Hara, J. Long, D. Horstmanshof, J. Chaffin, C. Elkins, P. Kanaly, E. Leiker, L. Gray, R. Dowling, S. Pagni, G. Bhat, P. Adkisson, S. Prabu, R. Sharma, S. Aggarwal, T. MacGillivray, A. Agnihotri, J. Madsen, G. Vlahakes, B. Rosengard, M. Semigran, S. Ennis, J. Camuso, R. Daly, S. Park, L. Durham, B. Edwards, C. Anderson, I. Penev, F. Arabia, P. DeValeria, E. Guyah, L. Lanza, R. Scott, E. Steidley, K. McAleer, T. Dewey, M. Magee, M. Mack, A. Anderson, T. Worley, D. Goldstein, S. Maybaum, D. D'Alessandro, N. McAllister, K. Brooks, D. Denofrio, D. Pham, H. Rastegar, A. Ehsan, H. Cote, M. Camacho, M. Zucker, L. McBride, S. Shah, C. Carr, R. Cecere, N. Giannetti, C. Barber, T. Icenogle, J. Everett, D. Sandler, M. Pulhman, J. Rich, J. Herre, L. Pine, K. Fleischer, M. McGrath, C. Klodell, J. Aranda, N. Staples, W. Dembitsky, B. Jaski, R. Adamson, S. Baradarian, S. Chillcott, A. Tector, B. Pisani, J. Crouch, F. Downey, D. Kress, M. McDonald, D. O'Hair, M. Savitt, M. Miller, C. Sheffield, C. Caldeira, L. DiChiara, V. Rao, J. MacIver, J. Kirklin, R. Bourge, D. McGiffin, S. Pamboukian, B. Rayburn, J. Tallaj, D. Baldwin, J. Cleveland, J. Lindenfeld, A. Brieke, B. Reece, S. Shakar, E. Wolfel, A. Cannon, B. Griffith, E. Feller, J. Brown, L. Romar, F. Pagani, K. Aaronson, J. Haft, T. Koelling, B. Dyke, E. Devaney, S. Wright, L. McGowan, A. Boyle, R. John, L. Joyce, M. Colvin-Adams, E. Missov, C. Toninato, R. Kormos, D. McNamara, K. Lockard, T. Massey, L. Chen, W. Hallinan, V. Chiodo,

P. Hobart, E. Verrier, D. Fishbein, C. Salerno, G. Aldea, S. Andrus, N. Mokadam, N. Edwards, M. Johnson, W. Kao, T. Kohmoto, J. Yakey, A. Li, S. Boyce, L. Miller, L. Sweet, K. Petro, M. Shah, L. Miller, F. Pagani, O. Frazier, S. Russell, Y. Naka, M. Slaughter, D. Farrar, C. Yancy, S. Hunt, W. Holman, W. Richenbacher, D. Heitjan, S. Moore, V. Jeevanandam, C. Thomas, S. Gordon, L. Damme, J. Heatley, and S. Reichenbach (2009). "Advanced heart failure treated with continuous-flow left ventricular assist device," *New England Journal of Medicine*, 361, 2241–2251.

Snowden, J. M., S. Rose, and K. M. Mortimer (2011). "Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique," *American Journal of Epidemiology*, 173, 731–738.

Squire, S. B., A. R. Ramsay, S. van den Hof, K. A. Millington, I. Langley, G. Bello, A. Kritski, A. Detjen, R. Thomson, F. Cobelens, and G. H. Mann (2011). "Making innovations accessible to the poor through implementation research," *Int. J. Tuberc. Lung Dis.*, 15, 862–870.

Stabile, G., F. Solimene, E. Bertaglia, V. L. Rocca, M. Accogli, A. Scaccia, N. Marrazzo, F. Zoppo, P. Turco, A. Iuliano, G. Shopova, C. Ciardiello, and A. D. Simone (2009). "Long-term outcomes of crt-pm versus crt-d recipients," *Pacing and Clinical Electrophysiology*, 32, S141–S145.

Stefanski, L. A. and D. D. Boos (2002). "The calculus of m-estimation," *The American Statistician*, 56, 29–38.

Stewart, G. C. and L. W. Stevenson (2011). "Keeping left ventricular assist device acceleration on track," *Circulation*, 123, 1559–1568.

Stroup, D. F., J. A. Berlin, S. C. Morton, I. Olkin, G. D. Williamson, D. Rennie, D. Moher, B. J. Becker, T. A. Sipe, and S. B. Thacker (2000). "Meta-analysis of observational studies in epidemiology: A proposal for reporting," *Journal of the American Medical Association*, 283, 2008–2012.

Stuart, E. A. (2010). "Matching methods for causal inference: A review and a look forward," *Statistical Science*, 25, 1–21.

Sutton, A. J. and J. P. T. Higgins (2008). "Recent developments in meta-analysis," *Statistics in Medicine*, 27, 625–650.

Tchetgen, E. T. (2014). "The control outcome calibration approach for causal inference with unobserved confounding," *American Journal of Epidemiology*, 175, 633–640.

Tierney, J. F., L. A. Stewart, D. Ghersi, S. Burdett, and M. R. Sydes (2007). "Practical methods for incorporating summary time-to-event data into meta-analysis‘," *Trials*, 8.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). "Super learner," *Statistical Applications in Genetics and Molecular Biology*, 6.

van der Laan, M. J. and J. M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*, New York: Springer.

van der Laan, M. J. and S. Rose (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York: Springer.

van der Laan, M. J. and D. B. Rubin (2006). "Targeted maximum likelihood learning," *The international journal of biostatistics*, 2.

Woods, B. S., N. Hawkins, and D. A. Scott (2010). "Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: A tutorial," *BMC Medical Research Methodology*, 10.